SUMMARY REPORT OF THE 3RD IAEA TECHNICAL MEETING ON FUSION DATA PROCESSING VALIDATION AND ANALYSIS (FDPVA)

D. Mazon
CEA, IRFM
Saint-Paul-lez-Durance, France
Email: Didier.MAZON@cea.fr

S. M. Gonzalez de Vicente
International Atomic Energy Agency, Department of Nuclear Sciences and Applications
Vienna, Austria
Email: S.M.Gonzalez-De-Vicente@iaea.org

M. Churchill
Princeton Plasma Physics Laboratory
Princeton, NJ, United States
Email: rchurchi@pppl.gov

A. Dinklage
Max Planck Institut fuer Plasmaphysik
Greifswald, Germany
Email: dinklage@ipp.mpg.de

R. Fischer
Max Planck Institut fuer Plasmaphysik
Garching, Germany
Email: rainer.fischer@ipp.mpg.de

M. Jakubowski
Max Planck Institut fuer Plasmaphysik
Garching, Germany
Email: marcin.jakubowski@ipp.mpg.de

A. Murari
EUROfusion Programme Management Unit, Culham Science Centre
Abingdon, United Kingdom
Email: Andrea.Murari@euro-fusion.org

M. Romanelli
UKAEA, Culham Science Centre
Abingdon, United Kingdom
Email: Michele.Romanelli@ukaea.uk

J. Vega
CIEMAT
Madrid, Spain
Email: jesus.vega@ciemat.es

G. Verdoolaege

Ghent University, Department of Applied Physics
Gent, Belgium
Email: geert.verdoolaege@ugent.be

M. Xu
Southwestern Institute of Physics
Chengdu, Sichuan, China
Email: minxu@swip.ac.cn

Abstract

The Third IAEA Technical Meeting on Fusion Data Processing, Validation and Analysis took place at the IAEA Headquarters in Vienna, Austria, from 28th May to 31st May 2019 and brought together more than 60 scientists and engineers from 19 Member States, the European Commission and ITER Organisation working on data analysis and machine learning methods for the processing of fusion data, collected either from experimental diagnostics or from plasma simulations.

“Accurate data processing leads to a better understanding of the physics related to nuclear fusion research. It is essential for the careful estimate of the error bars of the raw measurements and processed data,” said D. Mazon (Chair, CEA, France) in his introductory talk and progress has been shown in this direction during the meeting. In particular the meeting discussed new developments in fusion R&D applications in the following areas: inversion techniques, such as tomography; magnetic topology reconstruction, such as equilibrium reconstructions; system identification; scaling laws determination and their accuracy for extrapolation from current machines to fusion reactors; model-based algorithms for control applications; identification of spurious and undesired events, such as disruptive phenomena or hot spots in infra-red images using sophisticated mathematical techniques like neural network and support vector machines.

Discussions also focused on the potential use of these techniques for ITER, in particular, on their relevance to the first ITER plasma. The use of IMAS (Integrated Modelling & Analysis Suite) infrastructure - a convenient platform that allows implementation of different simulation codes in the same format - and synthetic diagnostics that could be coupled with this structure, in order to test the different mathematical approaches developed by the meeting participants, were explored in detail. Recent results and progress in the development of new tools will be discussed during the next meeting on the topic, scheduled in 2021.

Keywords: Integrated Data Analysis, Data Validation, Bayesian Techniques, Neural Networks, Machine Learning, Disruption Predictors, Image Processing.

1. INTRODUCTION

During the Third IAEA Technical Meeting on Fusion, Data Processing Validation and Analysis, 47 presentations were given by representatives from 23 institutions in 16 Member States. Sessions were devoted to the following topics: (1) Uncertainty propagation of experimental data in modelling codes; (2) Data analysis lessons learnt, best practices and proposals for ITER; (3) Regression analysis: profiles, scaling and surrogate models; (4)

Learning in non-stationary conditions for experimental design and predictions; (5) Inverse problems; (6) Image processing; (7) Causality detection in time series; (8) Synthetic diagnostics, integration, verification and validation; (9) Deep learning; and (10) a final session for discussion and conclusions that is also summarized in this report. In addition, a training course on “A brief overview of probability theory in data science" was provided on Monday 27th May by G. Verdoolaege (Ghent University, Belgium). All of the individual presentations are available online and can be accessed at the Technical Meeting website[2].

In this report, the technical papers are summarized in their presentation order within each session. The International Programme Advisory Committee members of the Technical Meeting were: D. Mazon (Chair, CEA, France), M. Churchill (Princeton Plasma Physics Laboratory, USA), A. Dinklage (IPP Greifswald, Germany), R. Fischer (IPP Garching, Germany), N. Howard (MIT, USA), S. Kajita (Nagoya University), D. McDonald (UKAEA, UK), A. Murari (EUROfusion Programme Management Unit, UK), J. Stillerman (MIT, USA), J. Vega (CIEMAT, Spain), G. Verdoolaege (Ghent University, Belgium), and M. Xu (SWIP, China). The IAEA Scientific Secretary of this meeting was Sehila M. Gonzalez de Vicente.

2. SESSION SUMMARIES

In this section we present a brief summary of the 9 sessions that covered specific topics of interest for the Fusion Data Processing Validation and Analysis, focusing in particular on the main highlights and progresses including outcome of the general discussions.

2.1 Uncertainty Propagation of Experimental Data in Modelling Codes

R. Fischer (IPP Garching, Germany) chaired this session on uncertainty propagation of experimental data in modelling codes.

Plasma modelling codes are essential for the interpretation of experiments and the design of operational scenarios for present-day fusion devices, as well as for the study of ITER plasma scenarios and for the design of DEMO. These codes rely on various measured input quantities and modelling assumptions which are subject to uncertainties. Therefore, the uncertainties of all input parameters have to be quantified in a suitable way to be used in modelling codes. Additionally, sophisticated modelling codes like SOLPS or GENE might be computationally expensive, depending on the system size and the completeness of the model used and, quite often, cannot be used for extensive case studies. Effective methods for uncertainty treatment to find the reliability of code results have to be developed. This session highlighted and discussed some of the major topics of uncertainty quantification (UQ), uncertainty propagation (UP) and model validation (V). The session started with an introduction to the concept of Integrated Data Analysis (IDA) where plasma parameters and their uncertainties as well as synergistic effects from all available measurements and modelling information in a rigorous Bayesian framework

[2]https://nucleus.iaea.org/sites/fusionportal/pages/3rd%20tm%20on%20fusion%20data%20processing/general-information.aspx

were described. Two methods of estimating electron and ion density, temperature and velocity profiles plus their logarithmic gradients and the uncertainty of equilibrium parameters, and, in particular, the uncertainty of the separatrix position were shown in a second presentation. A third presentation highlighted the challenges of UQ, UP and V for the plasma edge code SOLPS-ITER, followed by a presentation about advanced adjoint techniques for an effective sensitivity analysis and error propagation study for plasma edge codes. The last presentation proposed the method of the Unscented Transform to calculate in real-time (RT) plasma stability parameters from machine learning (ML) approximations of stability codes, including the measurement uncertainties.

2.2 Data Analysis Lessons Learnt, Best Practices and Proposals for ITER

D. Mazon (CEA, France) chaired this session on data analysis lessons learnt, best practices and proposals for ITER.

EUROfusion is currently utilising five different tokamaks (JET, TCV, AUG, MAST-U and WEST) and one stellarator (W7X) to carry out its research plan. Work in EUROfusion is highly collaborative and analysis of the data from the above devices by geographically dispersed scientists requires a high degree of standardization and the development of a common data platform. The EUROfusion project Code Development for Integrated Modelling (CD) aims to deliver a unified data system and standards for code integration along with workflows for data analysis, code verification and validation. Fusion data are stored around Europe in different file formats, often using different units and sign conventions. Therefore, the use of fusion data across tokamaks for integrated modelling or Artificial Intelligence (AI) applications requires building either translators for each data format or a common data structure. Both choices have pros and cons: translators are fast to adopt and do not require any effort from the data provider to adapt their data format to a common data structure, while a standard data structure requires commitment to learn and adopt a new set of conventions, rules and meanings. However, translators are only sufficient for 'formal situations' e.g. loose code coupling, sporadic exchange of information, and are not suitable for low level integration (strong code coupling). The adoption of a common data structure guarantees that all codes will receive the correct information at the deepest level.

In this framework, ITER leads 'the way' to collaborative development and data exchange via developing IMAS (the ITER Modelling and Analysis Suite) [1]. The backbone of the IMAS infrastructure is a standardized, machine-generic data architecture that represents simulated and experimental data with identical structures. IMAS includes a set of tools to access data and design integrated modelling workflows as well as first plasma simulators workflows. It is important to note that IMAS originates from the EU-ITM data model (CPO) [2]. The European activity of CD[1] stems from the pioneering work done by the European Task Force Integrated tokamak Modelling (ITM). The choice of CD is to fully embrace the IMAS and ITER data structure (IDS) for EUROfusion data standardization and code integration. A coordinated

[1]http://wpcd-workflows.github.io

activity started in 2019, involving data experts of all the EUROfusion tokamaks, to develop routines for the mapping of tokamak data in IMAS.

The principle vehicle adopted by EUROfusion for the mapping of experimental data in IMAS is UDA (Universal Data Access). UDA is presently used to map existing experimental and machine description data from different machines into IMAS data format. The mappings are machine specific and use different technologies: XML, Matlab, etc. A UDA server is present on the EUROfusion Gateway, a computer cluster dedicated to data analysis and integrated modelling, which allows to access data from all EUROfusion experiments and map them in IMAS files. The Gateway represents the 'virtual EUROfusion laboratory', where data from all EUROfusion tokamaks can be accessed in IMAS and high-level analysis workflows can be executed.

The coupling of physics codes ported in IMAS is achieved within CD by using the Kepler integration framework to deliver complex workflows. Three Kepler workflows are presently released to users and ready for validation on experimental data: the equilibrium reconstruction workflow (EQRECONSTRUCT) and MHD stability workflow (EQSTABIL), the European Transport Simulator (ETS) and the Edge Turbulence Workflow with synthetic diagnostics. Each of the above workflows features several codes all ported in IMAS.

During the meeting it was stressed that ITER's strategy for diagnostic data analysis is based upon the adoption of best practices and recommendations from today's machines. The aim is to extract the maximum possible information from the available diagnostic set at each phase of the ITER device's configuration as it evolves according to the Staged Approach described in the ITER Research Plan [3]. As is common on today's machines, a hierarchy of approaches will be taken for the processing of diagnostic signals, their interpretation and inference ranging from simple scaling using calibration parameters, through more sophisticated physics analysis including the generation of kinetically constrained equilibria and interpretive transport analyses, to a rigorous inference approach in which as few, but explicitly stated, prior assumptions as possible are made in order to have the most objective interpretation of the measurements as possible.
Whilst the detailed implementation of the process for combining diagnostic contributions and validating the measurements is still in its early stages, the need for a strictly managed, yet flexible, process is recognised by all participants during the meeting, and the infrastructure to support such activities is already maturing. This infrastructure is the Integrated Modelling & Analysis Suite (IMAS) previously described that is implemented using expertise and technologies developed within the ITER Members' research programmes. The unifying element of IMAS is its use of a standardized data model capable of describing both experimental and simulation data from any device. The inclusion of machine description data also allows the development and application of analysis software that not only works for any configuration of the ITER machine occurring within its lifetime, but also for other machines which adopts the same self-describing data model. This allows software to be extensively validated within the ITER Members' programmes before it is deployed on ITER.
It is also recognised during the discussions that a relatively significant investment of effort is needed to develop the necessary diagnostic data analysis workflows within IMAS. Here the ITER Organization looks for a strong collaboration with the data analysis community to advise on the best approaches to adopt, to provide example analysis workflows, ideally in IMAS, and in the longer term, to apply, validate and improve the workflows developed for eventual deployment on ITER.

Nowadays, processing all information of a fusion database is a much more important issue than acquiring data. Fusion devices produce tens of thousands of discharges but only a very limited part of the collected information is analysed (physics studies are normally limited to a few tens of shots). Due to the fact that diagnostics produce the same morphological patterns in the signals for reproducible plasma behaviours, these plasma behaviours are recognised in experimental signals by the identification of known patterns. Therefore, the analysis of physical events requires their identification and temporal location in the signals. It should be noted that the recognition and location of plasma events are the main concerns in relation to the analysis. These searching processes are usually manual, complex and very time consuming. But taking into account that long pulse devices, such as W7X and ITER, will have databases with very large number of signals and very long records, relevant temporal segments in signals have to be found in an automatic way. It should be mentioned that 'relevant' means 'with interest from some point of view: either physics or machine control'.

During the meeting it was emphasized that databases in nuclear fusion are characterised by a large variety of signals and data representations: 2D (for example, time series), 3D (for instance, quantities with both spatial and temporal resolution) and video-movies (infrared and visible cameras). Nowadays, big data techniques have shown maturity enough in multiple domains and can be essential candidates for data analysis in high temperature plasmas. The term 'big data' has to be understood as a set of machine learning methods (classification, regression, clustering and dimensionality reduction) to process big amounts of complex data whose size does not allow a standard analysis with typical database software tools. The application of big data techniques in nuclear fusion comprises four objectives: data retrieval, searching for patterns, creation of data-driven models (to explain particular aspects of the plasma nature) and anomaly detection.

During the FDPVA meeting some presentation focussed on anomaly detection. A quiet plasma evolution usually generates stationary signals whose variations can be explained by signal noise. However, abrupt changes or the sudden presence of periodicities in the signals generally reveal modifications in the plasma evolution. These modifications can correspond to either known events (for example, heating power injection, confinement transitions, incoming disruptions or gas injection) or off-normal plasma behaviours. The automatic temporal location of abnormal plasma behaviours is of big interest to determine the set of signals that show variations around the same time interval and to analyse the possible causes. Moreover, data-driven models can be generated to predict and explain these particular plasma behaviours. As an example, a particular algorithm was presented for off-line recognition of relevant events and to find relationships among them. The algorithm is a six-step process:

Step 1: to define a dataset of signals and a range of discharges.

Step 2: to determine times in each discharge when individual signals show anomalies. There are several automatic methods to carry out this. Some of them are: detection of outliers through a generalised linear regression model, use of martingales for testing exchangeability, the Universal Multi-Event Locator (UMEL) technique, spectral clustering optimization, analysis of the temporal evolution of the Fourier components of a signal and the use of deep learning methods.

Step 3: to choose morphological patterns within a time interval tS around each anomaly time in every discharge.

Step 4: to define multi-signal patterns (MSP). A MSP is characterised by the morphological patterns of all the signals with a common time interval. A criterion to define the common time interval is necessary taking into account all MSPs in all discharges of the dataset.

Step 5: to group the MSPs into a number of sensible clusters in an unsupervised way (this reveals the organization of the MSPs).

Step 6: to develop supervised classifiers with the classes of step 5. This supervised classifiers can be implemented under real-time conditions.

The global nature of the ITER project, along with its projected ~petabyte per day data generation, presents a unique challenge, but also an opportunity for the fusion community to rethink, optimize and enhance our scientific discovery process. Recognizing this, collaborative research with computational scientists was undertaken over the past several years to create a framework for large-scale data movement across wide-area networks (WANs), to enable global near-real time analysis of fusion data. This would broaden the available computational resources for analysis/simulation and increase the number of researchers actively participating in experiments.

Such a federated analysis framework would enable faster and better analysis, leading to better decision making and, therefore, increased scientific productivity for ITER and other fusion experiments. A wide variety of analyses would be enabled, including anomaly detection of the various diagnostic outputs, integrating various diagnostics for model comparison and physics extraction and ML to identify phenomena of interest.

2.3 Regression Analysis: Profiles, Scaling and Surrogate Models

G. Verdoolaege (Ghent University, Belgium) chaired this session on regression analysis: profiles, scaling and surrogate models.

The goal of this session was to highlight the ubiquitous role of regression analysis in various methods and applications in fusion data analysis. The most common application is the fitting of a parametric function (a curve or surface) to experimental data, to capture a general trend of a response variable, when varying one or multiple predictor variables. In fusion, fitting of (radial) plasma profiles is another important application, nowadays handled either parametrically (e.g. using splines) or in a nonparametric way (e.g. using a Gaussian process). Recently, several applications of surrogate modelling have become of great interest in fusion, often employing neural networks to emulate codes that model certain aspects of plasma behaviour. Various methods and tools exist and are being developed to carry out these tasks by some form of regression analysis. Many techniques originate from the classical field of statistics, while others have been developed within the ML community, both for the purpose of interpolation and prediction. In this session a broad variety of techniques and applications of regression analysis in fusion were covered.

Real-time control is needed to steer the plasma away from unstable magnetic equilibria for reliable routine operation of a tokamak plasma. This requires fully automated real-time operation of diagnostics, processing of the raw measurements and operation of actuators on a time scale that is sufficiently small to allow control of MHD activity and transport.

To illustrate the role of machine learning in interpreting diagnostic measurements for predicting events that may adversely affect the equilibrium, prediction of disruptions and tearing modes was carried out at DIII-D, comparing several ensemble learning methods. Furthermore, prediction of plasma profile evolution was discussed, along with control of suitable actuators for pursuing the most desired plasma profile.
The approach was demonstrated by training a recurrent neural network to predict various plasma profiles (electron and ion density and temperature, rotation, q-profile, etc.) at DIII-D.

The session also contained a short talk related to confinement regime identification and profile evolution prediction. Herein, a classification scheme was presented for identification of confinement regimes in Alcator C-Mod. Excellent performances were obtained in classification of L-, H- and I-mode plasmas, using a neural network trained with global plasma parameters.

Gaussian process regression (GPR) for profile fitting was discussed. This technique allows for nonparametric fitting of plasma profiles in one (e.g. radial) or multiple dimensions. This means that no specific functional form or fixed basis functions are used to model the profile, instead relying on regularization of the fit by imposing a certain degree of correlation between neighbouring profile points. As a result, the fit is flexible and is able to adapt to the required level of smoothness and feature resolution. The output comes in the form of a multivariate normal distribution, hence providing error bars on the fitted profile. Moreover, spatial derivatives of the fitted profile can be calculated easily and reliably, again including error bars.

The field of surrogate modelling for turbulent transport codes was then discussed in depth. The research stems from the important motivation for experimental design, optimization of future devices, and real-time control of plasma parameters based on high-level physics modelling. The typical computational requirements of transport codes so far have limited their use in large-scale applications. However, emulation of such codes by means of surrogate models is becoming a feasible alternative to accomplish these goals. This opens a whole new area of applications of integrated modeling and even full-device modelling for prediction, control and uncertainty quantification.

Fusion scaling laws and the techniques that can be used to estimate them were presented. First, a case was made for data-driven approaches to theory development for characterizing complex physical systems, such as fusion plasmas. In part, this is motivated by the huge quantities of data gathered in fusion experiments, with various complex, nonlinear interdependencies, and the recent proliferation of powerful pattern recognition techniques that can be used to characterize these dependencies. In a way, the model is "extracted" from the data by machine learning techniques. The technique used here to accomplish this is symbolic regression, which involves combinations of mathematical operators to construct an analytical model for the data. This is an optimization problem that can be solved using genetic algorithms, which themselves rely on evolutionary principles to find the most suitable (the "fittest") model. The method is able to construct physically meaningful dimensionless quantities from measurements of the system.

Then, work of the Database Working Group (within the ITPA Topical Group on Transport and Confinement) on scaling laws, aimed at a revision of the IPB98 global H-mode energy confinement scaling. Some of the regression techniques that are being developed and applied to a recent version of the database that includes new data from JET with the ITER-like wall and ASDEX Upgrade with the full tungsten wall were shown.

The need for advanced regression techniques is motivated by the complex structure of the data. A robust Bayesian method and a new minimum distance technique one, called geodesic least squares regression (GLS), were applied and compared with the standard weighted least squares approach [4]. The robustness property of both methods stems from the model uncertainty which they take into account: the conditional probability distribution of the response variable contains extra nuisance parameters that are learned from the data.

Compared with IPB98(y,2), a weaker dependence of the confinement scaling on toroidal field and density is obtained, as well as a noticeable influence of plasma triangularity [5].

Predictions for ITER are somewhat lower than earlier results, but further work will be needed to allow more definitive statements regarding parametric dependencies and predictions of global H-mode confinement.

Numerous and very diverse applications exist for regression analysis applied to data from fusion diagnostics for model fitting, prediction, model validation, etc. For the practitioner, it is important to analyse the requirements of the problem and the application at hand, and to consider the characteristics of the data, before choosing a specific regression method from the extensive fields of statistics and ML.

It is therefore essential for the fusion community to establish a toolset of adequate and well-implemented regression techniques, which can be applied without the need for specialized statistical expertise (this goes not only for regression, but also for other data analysis tasks).

In addition, whereas many regression models, like power laws used for scaling analysis, do not have a clear physical motivation, the data analyst can in principle incorporate any type of physical knowledge in the model or in the fitting constraints (e.g. prior probability). Thus, the purely “data-driven” and “physics-based” paradigms are merely two extremes of a continuous spectrum of model fitting approaches.

2.4 Learning in Non-Stationary Conditions for Experimental Design and Predictions

A. Murari (EUROfusion Programme Management Unit, UK) chaired this session on regression analysis: profiles, scaling and surrogate models.

Machine learning methods use sets of data for training/test and the resulting models are applied to real situations. The main assumption which was extensively discussed among the participants in ML developments is the hypothesis of independent and identically distributed data (“iid” hypothesis). This means that the training, test and final application data follow the same (but unknown) probability distribution function. However, real systems can violate the “iid” assumption and several data distributions can appear not only in training and test data but also in the real-life deployments and experiments.
From the many applications of ML systems in situations violating the “iid” hypothesis, three examples are considered in this sector: planning of new experiments to explore new regions of the operational space, adaptive methods from scratch to predict disruptions and, finally, anomaly detection techniques for both disruption prediction and predictive maintenance. The planning of new experiments to explore new regions of the operational space means that no data for the new regions are available. Therefore, the examples of known regions are used for training and to gain knowledge about the unexplored parts of the operational space.

A second application of ML methods to systems, which violate iid assumption, is adaptive learning based on ensembles of classifiers for disruption prediction. This subject is very relevant to ITER. Thermonuclear plasmas are physical systems that usually present memory and historical effects and, therefore, the “iid” assumption is not satisfied. However, ML methods are used to predict or recognise plasma behaviours such as L/H transitions or disruptions. Focusing the attention on disruptions, it is important to mention that ML methods are used to predict disruptions due to the need to counteract the following weaknesses: lack of reliable theoretical models, impossibility of collecting many examples in the next generation of tokamaks (the proposed tools implement an approach from scratch) and rapid variations of the experimental conditions.

It is important to mention the potential relevance of predictive maintenance in fusion devices using anomaly detection techniques. One possible technique for anomaly detection can be dimensionality reduction to two components by means of principal components analysis. Under normal conditions, points form a compact cluster. In case of off-normal conditions, points appear ‘far’ from the cluster. The term ‘far’ implies the computation of a distance to the cluster centre. To take into account the covariance in the data, the Mahalanobis distance can be used and a threshold can be determined to identify off-normal situations. This technique is completely equivalent to the one implemented in the JET real-time network to recognise disruptive behaviours [6]. A second technique for anomaly detection which was also presented during the session is the use of an auto-encoder network trained on healthy data, in order to determine the mean absolute error input-output. It is important to note that no assumption about the distribution functions of the data is required.

Simple and efficient methods for disruption prediction were then presented during the session. Classification and Regression Trees (CART). CART is a supervised methodology that recursively allows partitioning the data space. It results in a series of ‘propositional logic’ rules that can be represented graphically by a decision tree. However, CART trees are not very stable. Small changes in the training set can result in major differences in the final trees and, therefore, in the final classification. To counteract this fact, the trick is to increase diversity by training with slightly different sets. In this way, ‘weak’ learners are developed and then they can be pooled together to create a ‘strong’ ensemble classifier.

Finally, it was mentioned the potential relevance of predictive maintenance in fusion devices using anomaly detection techniques. One possible technique for anomaly detection can be dimensionality reduction to two components by means of principal components analysis. Under normal conditions, points form a compact cluster. In case of off-normal conditions, points appear ‘far’ from the cluster. The term ‘far’ implies the computation of a distance to the cluster centre. To take into account the covariance in the data, the Mahalanobis distance can be used and a threshold can be determined to identify off-normal situations. A second technique for anomaly detection is the use of an auto-encoder network trained on healthy data in order to determine the mean absolute error input-output. It is important to note that no assumption about the distribution functions of the data is required.

2.5 Inverse Problems

M. Xu (SWIP, China) was the chair of this session on inverse problems in fusion science.

Inverse problems are some of the most important mathematical challenges in science and mathematics, which involves parameters that we cannot directly observe.

The equilibrium reconstruction is among the most crucial inverse problems in the fusion community, which provides essential magnetic geometry, current and pressure profiles information necessary for tokamak operation and data analysis. First, the motional Stark splitting of the spectral lines (MSE-LS) reconstruction algorithm has been successfully incorporated with EFIT response matrix [7] and proved its potential for ITER application. Second, based on EFIT, a complementary parallel computing approach has been implemented, namely P-EFIT. This algorithm has been tested and validated both on DIII-D and EAST backgrounds, demonstrating the promising real-time ability by using GPU. Third, model-assisted full kinetic equilibrium reconstruction pressure and current profiles based on several different theory models has been discussed. The theory-based constraints provide good guidance and self-consistency to equilibrium reconstruction. Finally, a strategy of using extended MHD linear simulation to guide 3D perturbed equilibrium reconstruction was proposed [8].

Calibration of Magnetic Detectors on MAST-U discussed the specific inverse problem of the calibration of magnetic detectors [9]. This work was proposed as an innovative solution by seeking an optimum configuration for the meta-data using principles of Bayesian Inference.

Real-time aspect of full kinetic equilibrium reconstruction (KER), was made using LIUQE-RAPTOR coupling'. The present approaches for KER were reviewed and introduced in detail. Specifically, KER has been demonstrated in real-time for the first time, where the application was delivered by coupling free boundary equilibrium reconstruction (LIUQE) [10] and radial transport code (RAPTOR) [11-12].

Progress on applications of Position Sensitive Detector (PSD) And applications of 1D PSD has been demonstrated on neutron tomography purposes [13] and 2D PSDs are widely used in small angle neutron scattering facilities [14].

2.6 Image Processing

M. Jakubowski (IPP Greifswald, Germany) was the chair of this session on Image Processing.

In magnetic fusion devices imaging, meant here as a process of obtaining images to secure information on a physics phenomenon, became a standard, widely used tool for machine protection, research and plasma control. The range of applications is very broad and includes, among others, real-time protection of plasma facing components with infrared cameras and spectroscopic detection of plasma flows or automatic offline edge filaments detection. The information content delivered by cameras can be very different, depending on the experimental conditions and wavelength. Additionally, presently used cameras can deliver up to hundreds of kiloframes per second, which poses a significant challenge for data analysis. Therefore, very different image processing and data handling tools need to be developed, in order to fully exploit the benefits of imaging diagnostics.

The main challenges to image processing can be divided into two categories:

- Extract required information in real-time, e.g. for machine protection. Here the main challenge is to automatically detect events and assess the risk their pose to a magnetic fusion device. A classification of detected events may be very wide and their definition ambiguous. Additionally, the general appearance of the frames (i.e. the level of noise, scene luminosity) may vary dramatically from experiment to experiment. High reliability and low number of false positives requires sophisticated methods for detection and semantic analysis of the images.
- Automatically process large amount of data, e.g. a database of images at Wendelstein 7-X after a few campaigns will contain petabytes of data. Analysing such amounts of data manually is impossible. New methods of information retrieval, event detection based on structural pattern recognition and sorting, need to be developed or adapted from different fields.

These challenges were discussed extensively during the session on image processing. New methods for image preparation, real-time processing and offline analysis of existing data were presented during the meeting.

2.7 Causality Detection in Time Series

J. Vega (CIEMAT, Spain) was the chair of this session on Causality Detection in Time Series

Longitudinal data analysis and the investigation of time series are basic elements of research in Magnetic Confinement Nuclear Fusion (MCNF). Indeed, the vast majority of diagnostic systems produce time series, which have to be carefully analysed to obtain robust conclusions. The relation between the quantities measured by these time series is often highly non-linear. Their interpretation is also complicated by the fact that the environment is very noisy and that MCNF plasmas are difficult to access for measurements. Reliably determining causal relations between these signals is therefore a quite challenging task.

Within the community, by far the most used statistics to obtain information about time series are covariance and correlation coefficients. Two important limitations of these tools are the fact that they detect only the linear relations between the variables and do not depend on the order of the signals; therefore they are incapable of answering questions about directionality. This situation is particularly unsatisfactory because, since the 60ies, much effort has been applied in other fields to develop more advanced statistical techniques to determine the causal relations in longitudinal data. This resurgence of interest for causality was triggered by the introduction of Granger Causality (GC) by Clive Granger [15]. The approach originated in econometrics and is based on the concept of prediction error, proposed first by N. Wiener in 1956. The main consideration is that if a variable x(t) is an appreciable cause of y(t), in the sense that it has a detectable effect on y(t), it should be helpful in predicting y(t). In other words, the information contained in the past of x(t) should improve the prediction of the future of y(t). GC testing has become very popular recently, due to its straightforward implementation. In the last decades additional techniques have been developed to complement and improve GC.

One important recent generalization of GC is an information theoretic criterion called Transfer Entropy (TE). For example, the application of the transfer entropy might consist of quantifying the causal impact of one (turbulent) variable on another. TE is independent from signal amplitude or waveform and is directional ($Y \rightarrow X$), unlike traditional techniques such as the correlation. This property makes it possible to track the propagation of small perturbations in a turbulent 'sea' of fluctuations, which is very useful for transport studies.

A first case study is offered by a so-called 'slow' L-H transition at TJ-II that was presented in this session as well as an analogous study done at W7-X using ECE measurements. At JET, the high resolution ECE system was used to study the propagation of heat pulses due to sawtooth crashes too.

The vast majority of tokamaks exhibit also relaxation oscillations in the centre of the plasma in many experimental conditions. The signature of these relaxations is a non-linear oscillation of the central plasma parameters, which is similar to a sequence of saw-teeth, from which they derive their name. If the saw-teeth become too large, their crashes can trigger very dangerous instabilities, such as neo-classical tearing modes (NTM), associated with flux surfaces, where the safety factor q is a rational number (typically q=2 or q=3/2). Such NTM, in addition to degrading the confinement properties of the plasma, can even trigger disruptions. In the next generation of machines, it will therefore be extremely important to have available some form of control on the saw-teeth frequency, so that their amplitude can be kept at manageable levels.

2.8 Synthetic Diagnostics, Integration, Verification and Validation

A. Dinklage (IPP Greifswald, Germany) was the chair of this session on Synthetic Diagnostics, Integration, Verification and Validation.

How much information is needed to control fusion plasmas? How much can virtual instruments support the development and implementation of plasma diagnostics? What are the critical margins for sensors to ensure safe operation? How much does the simulation of plasmas save valuable machine time and can contribute to accelerate the scientific program of large-scale fusion devices? Being subject of current research, the fusion community converges to the opinion that simulations of measurements has a high potential to contribute to answering those questions. Virtual instruments appear to be highly valuable for the preparation of future plasma devices like ITER and DEMO. The session on synthetic diagnostics, application and validation provided a state-of-the-art revision of the field and identified future directions and collaboration potentials.

A specific view reflecting the relevance of diagnostics simulations was pointed out. Virtual instruments are considered an integral part of the conceptual approach for experiment preparation and conduction on ITER. Specific consideration in real-time control for stable, high-performance operation were addressed as well. A key element for advanced real-time control is to include physics model-based simulations. As one example, finite state-modelling for the control of NTMs was demonstrated on TCV for proof-of-principle: a discrete state supervised controller could achieve tearing mode stabilization and beta control at high heating power. How much virtual instruments can support diagnostics developments was exemplified. While commercial ray tracing software does not account for polarization effects, it is exactly the propagation of polarized light that matters in fusion diagnostics. Examples are the Motional Stark Effect or phase contrast imaging. Forward functions are also used in (Bayesian) data analysis. Good data interpretation consequently serves for the validation of the virtual instruments. An example taking benefit from multiple diagnostics analyses was presented for the analysis of Thomson scattering data and interferometry data from the Large Helical Device.

The benefits of a concise combination of diagnostics data for divertor studies were outlined. Virtual Bayesian multi-diagnostics inference were used to assess how physics questions for the super-x divertor being installed on MAST-U can be resolved. Simulations with synthetic instruments from line-integrated measurements (filtered cameras) combined with point measurements (Thomson scattering, Langmuir probes) were done with input from predictive SOL modelling (from SOLPS). It was concluded that the reconstruction of plasma properties is expected at useful accuracy. A different aspect employing Bayesian techniques addressed the quantification of uncertainties of combined data sets in impurity transport studies. A Bayesian framework was developed at MIT to derive impurity transport coefficients. Investigations were made for ELM-free stationary, high performance discharges on Alcator C-Mod. Validation of toroidal rotation and ion temperature measurements was reported. The report discussed studies on KSTAR and EAST. Cross comparison between (XICS) and charge-exchange spectroscopy add an additional important aspect for the specific validation of data analysis techniques. The results showed good agreement of the measurements.

The cross comparison of spectroscopic measurements and probe data was addressed as well. Measurements on the GLAST-II and III spherical tokamaks will incorporate successively different electrostatic probes and spectrometers. Fundamental to quantitative assessments, however, is the proper quantification of uncertainties. An example of plasma wall interaction analysis was provided. Uncertainty quantification for computer models need to catch a large variety of uncertain model parameters, e.g. in sputtering yields.

Simulations of measurements are getting more and more into the focus of current fusion research. Benefits in the design and implementation of plasma diagnostics have been revealed and were reported to lead to more efficient diagnostics. The value of virtual instruments was agreed to have beneficial prospects for the conduction of experiments on large devices. Since comprehensive flight-simulator-like discharge preparation accelerates the planning of experiments and a prediction of coverage of measurements, a suite of validated synthetic diagnostics appears to be of large benefit to large fusion devices. In order to obtain tools of high maturity, synthetic diagnostics can be validated on existing devices. Key aspects are UQ, multi-diagnostics approaches and validation of measurements. Physics models for simulating measurements and for use in control applications are the main focus of activities. Discussions led to proposals to implement required activities in international networks, namely the ITPA groups.

2.9 Deep Learning

M. Churchill (Princeton Plasma Physics Laboratory, USA) was the chair of this session on Deep Learning.

This session contained presentations featuring the use of deep neural networks and other ML algorithms in a variety of applications within magnetic fusion research, summarised as follows.

A novel convolutional neural network architecture with dilated convolutions enables making predictions with raw, time-series sequences from fusion diagnostics which have long-range dependencies. Deep learning was discussed in the paradigm of viewing hidden layers as “filters”, which the neural network “learns”, in order to accomplish a specified task. This contrasts the usual analysis where humans define the filters or transforms needed for a task. Deep learning allows working directly with high dimensional data versus low dimensional data derived by models from diagnostic data.

One of the main topics of discussion centred on interpretability, understanding why the neural network was making the prediction of disruption or not. Currently no work by the authors has been dedicated to understanding the why, but it was discussed that interpretability algorithms exist for neural networks, which could be useful in explaining the decisions made. Another point of discussion was, without interpretability, how it is possible to extend the approach to new devices such as ITER. There would need to be research into transfer learning, the ability to transfer the neural networks learned on current devices and/or simulation and re-training these networks on smaller datasets from ITER. Also, matching between different diagnostics would be necessary, requiring synthetic diagnostics and simulation.

A second topic addressed during this session was disruption predictions using random forest feature contribution analysis. The availability of a huge amount of experimental data has enabled plasma researchers to overcome the limits dictated by a partial understanding of disruption physics by adopting data-driven techniques that use elaborate algorithms for disruption detection and prediction. Nevertheless, the goal of disruption research is not only to provide predictions of impending disruptions with hundreds of milliseconds of warning time but also to inform the Plasma Control System (PCS) on the offending feature(s), in order to steer the plasma away from the disruptive boundary and its deleterious consequences for the device. A ML-based Disruption Predictor using Random Forest (DPRF) has been recently developed for several different tokamaks. It has the main advantage of guaranteeing explainable predictions.

Real-time control and estimation in tokamaks with ML accelerated predictive models was presented. Present day and next step tokamaks will require precise control of plasma conditions, including the spatial distribution of rotation and current profiles, in order to optimize performance and avoid physics and operational constraints. This motivates expanding the availability of diagnostics in real-time as well as developing physics-model-based approaches to real-time plasma condition estimation, feedback control and scenario forecasting.

The session included discussion of neural network accelerated modelling of ITER Plasmas Compatible with IMAS. An integrated modelling workflow capable of finding the steady-state solution with self-consistent core transport, pedestal structure, current profile and plasma equilibrium physics has been developed and tested against DIII-D discharges. Key features of the achieved core/pedestal coupled workflow are its ability to account for the transport of impurities in the plasma, as well as its use of machine-learning-accelerated models for the pedestal structure and for the turbulent transport physics.

A new method, based on AI/Deep Learning, to forecast disruptions and extend considerably the capabilities of previous strategies, such as first principles-based and classical/shallow machine-learning approaches, as applied to large burning plasma systems such as the multi-billion-dollar international ITER project currently under construction. The key goal for ITER is to deliver fusion reactions capable of producing more power from fusion than is injected to heat the plasma and avoidance of tokamak disruptions, enabled through such predictions, is a pressing challenge.

Deep learning for plasma tomography and disruption prediction was also included in the session. This work presented two applications of deep learning to the processing of bolometer data at JET. In the first application, a Convolutional Neural Network (CNN) was developed (actually, the inverse of a CNN, sometimes called a deconvolutional network) to significantly accelerate the tomographic reconstruction of the plasma radiation profile. The reconstruction is much faster (from minutes to milliseconds), so it becomes possible to analyse the plasma profile at each time step across an entire pulse.

3 CONCLUSIONS

In the closing session of the meeting, the discussion started with a summary of the general impression about the meeting. In particular, advancement and significant maturation of activity and outcomes since the 2nd edition, organized in Boston, MA, USA in 2017, was noted. This is evident from a growing number of participants from an increasingly broader range of geographical areas, but also from a steadily widening range of topics, concerning both methods and applications.

As a general conclusion, a strong consensus was reached at the 3rd IAEA TM FDPVA for the involved community to take actions in a coordinated and results-oriented way toward the following goals:

- Take concrete steps to transfer key technologies from the domain of fusion data processing, validation and analysis to ITER, providing expertise regarding techniques and their practical software implementation. It was proposed to collect, in close collaboration with ITER specialists, the requirements for ITER data analysis and to identify priority areas of expertise and practices where the FDPVA community can contribute. This may lead toward a design review of ITER data analysis.
- Make an inventory of priority synthetic diagnostics for ITER, to be stored in a central repository with access for the fusion community. Transfer existing synthetic diagnostics to ITER, ensuring compatibility with IMAS. Identify possible contributors of additional synthetic diagnostics or additional modules in existing synthetic diagnostics.
- Initiate an integrated Bayesian workflow for a set of key ITER diagnostics, by means of an open, transferable and modular software framework that is compatible with IMAS.
- Take steps to increase validation of mature technologies for data processing, validation and analysis throughout the fusion community, e.g. by transferring tools developed at one device to others.
- It was stressed that the availability of sufficient manpower is key to realizing some of these goals. In addition, it is envisaged that the ITPA will play a central role in coordinating these actions, notably through the recently established Real Time Specialist Working Group within the ITPA Diagnostics Topical Group.
- Finally, the attendees of the IAEA TM FDPVA warmly welcomed the proposal by the Chinese delegation from SWIP to organize the next edition of the workshop in China in 2021.

4 ACKNOWLEDGEMENTS

The views and opinions expressed herein do not necessarily reflect those of the IAEA, European Commission or other organization. This work has been partially carried out within the framework of the EUROfusion Consortium and has received funding from the Euratom research and training programme 2014-2018 and 2019-2020 under grant agreement No 633053.

REFERENCES

[1] F. Imbeaux et al. 2015 Nucl. Fusion 55 123006

[2] A. Becoulet et al. Computer Physics Communications, Volume 177, Issues 1–2, July 2007, Pages 55-59

[3] ITER Research Plan within the Staged Approach: https://www.iter.org/technical-reports

[4] G. Verdoolaege et al., Nucl. Fusion 55, 113019, 2015.

[5] G. Verdoolaege et al., “First analysis of the updated ITPA global H-mode confinement database" 27th IAEA Fusion Energy Conference, EX/P7-1, Gandhinagar, India

[6] S. Esquembri et al. IEEE Transactions on Nuclear Science, 65(2): 836–842, 2018.

[7] L.L. Lao et al. 1985 Nucl. Fusion 25 1421

[8] N.M. Ferraro et al. 2013 Nucl. Fusion 53 073042

[9] J. M. Moret et al. Rev. Sci. Instrum. 69, 2333 (1998).

[10] J. M. Moret et al., Fusion Engineering and Design 91, 1-15 (2015).

[11] F. Felici et al. 2018 Nucl. Fusion 58 096006

[12] F. Carpanese et al. 2020 Nucl. Fusion 60 066020

[13] N. Abidin Ashari, MSc thesis 2009, University Sains Malaysia: http://eprints.usm.my/10231/

[14] A. A. Mohamed et al 2013 IOP Conf. Ser.: Earth Environ. Sci. 16 012141

[15] C. W. J. Granger, Econometrica Vol. 37, No. 3 (Aug., 1969), pp. 424-438