



Politechnika Łódzka

Instytut Informatyki

PRACA DYPLOMOWA MAGISTERSKA

Hybrid neural networks for anomaly detection in cyber-physical systems

Wydział Fizyki Technicznej, Informatyki i Matematyki Stosowanej

Promotor: dr hab. inż. Aneta Poniszewska-Marańda

Diplomant: inż. Ramzi Hadrich

Nr albumu: 227488

Kierunek: Informatyka

Specjalność: Computer Science & Information Technology

Łódź, 11.09.2020r.



Instytut Informatyki

90-924 Łódź, ul. Wólczańska 215, budynek B9

tel. 042 631 27 97, 042 632 97 57, fax 042 630 34 14 email: office@ics.p.lodz.pl

Abstract

Nowadays cyber-physical systems are widely used in different application domains. In parallel, machine learning algorithms are used widely to detect the anomalies in the behaviour of these systems. However, this detection is limited to two states: normal behaviour and faulty functioning. This master thesis aims to extend this detection to differentiate between attacks and normal faults. In first place, a power system is described as an example to work on. Then, various machine learning algorithms are evaluated on the given datasets, and this using two machine learning toolkits - scikit-learn and Weka. Later, various tools for feature analysis are presented and an algorithm to find the features that contributed the most into the false predictions is described. Finally, three solutions to the initial problem are presented and evaluated.

Keywords: machine learning, cyber-physical systems, anomaly detection, random forest, power system

Streszczenie

Obecnie systemy cyber-fizyczne są szeroko wykorzystywane w wielu różnych domenach. Jednocześnie, algorytmy uczenia maszynowego są szeroko stosowane do wykrywania anomalii w działaniu tych systemów. Ale ta detekcja jest ograniczona do dwóch stanów: normalne działanie i naturalne błędy. Ta praca magisterska ma na celu rozszerzenie rozróżniania tych stanów, aby wziąć pod uwagę również ataki na system. W pierwszej kolejności, system energetyczny został opisany jako przykład systemu cyber-fizycznego do dalszych rozważań. Następnie, dokonano oceny różnych algorytmów uczenia maszynowego na bazie zbioru danych z przedstawionego systemu. Do tego celu wykorzystano dwa narzędzia do uczenia maszynowego - scikit-learn i Weka. Później, omówione zostały różne narzędzia do analizy cech systemu. Oprócz tego, zaprezentowano algorytm umożliwiający ustalenie cech najbardziej wpływających na złą klasyfikację danych. Na koniec, trzy rozwiązania pierwotnego problemu zostały przedstawione i sprawdzone.

Słowa kluczowe: uczenie maszynowe, system cyber-fizyczny, detekcja anomalii, lasy losowe, system energetyczny

Preface and acknowledgements

This thesis was written in the framework of my 5-month internship at LORIA (Laboratoire Lorrain de Recherche en Informatique et ses Applications) in Nancy, France, within my Erasmus+ exchange, from September 2019 to July 2020, at Ecole Nationale Supérieure d'Electricité et de Mécanique (ENSEM), which is a part of University of Lorraine.

Despite COVID-19 pandemic, it was possible to continue and finish the work from home. Mr Abdelkader Lahmadi and Mr Jérôme François were there for me during the whole period of the internship. I wanted to say thank you to both of them, for the continuous help, which made this thesis a reality. I wanted also to thank Mr Lahmadi for proposing me this research project.

I am also thankful to the international office team at my home university as well as to my supervisor Mrs Aneta Poniszewska-Marańda for making this Erasmus+ exchange possible.

Finally, I wanted to thank also my family and friends for the support during the writing process of this thesis.

Contents

Introduction	6
1 Power system as a CPS example	10
2 Machine learning algorithms comparison	16
2.1 Machine learning toolkits	16
2.1.1 Weka	16
2.1.2 scikit-learn	17
2.2 Used machine learning algorithms	17
2.2.1 Random Forest	17
2.2.2 SVM	19
2.2.3 NaïveBayes	21
2.2.4 Multilayer perceptron	21
2.3 Metrics for classifiers comparison	23
2.4 Comparison results	25
2.5 scikit-learn further methods' analysis	34
3 Features' importance	45
3.1 Result's interpreters' comparison	45
3.1.1 LIME	45
3.1.2 ELI5	47
3.1.3 YellowBrick	47
3.1.4 Treeinterpreter	47
3.1.5 scikit-learn export_graphviz	47
3.1.6 dtreeviz	48
3.1.7 Summary	48
3.2 Features' importance determination	49

4 Model enhancement	53
4.1 Features values modification	53
4.2 Distance between features	56
4.3 Hidden Markov Model	58
Conclusions	61
List of Figures	63
List of Tables	65
List of Listings	66
Bibliography	67

Introduction

Cyber-Physical Systems (CPS) are nowadays widely used in different application domains, such as smart-homes, smart-cities, hospitals, etc... They are mainly composed of two entities: a cyber part consisting in a computing and networking component, and a physical part consisting in different controllers and sensors. The existence of a connected cyber part implies its susceptibility to multiple cyber threats. The malfunctioning of these systems, due to a cyber threat, can cause severe impacts on the real life and the safety of the community, for example a blackout or water contamination. That is why many algorithms have been designed for the security monitoring of those systems, in particular the anomaly and attack detection.

Nowadays, machine and deep learning algorithms are used to detect those anomalies and intrusions. But, in majority, they rely only on the cyber part of the systems and on the data describing their behaviour, ignoring their physical models. The **goal** of this master thesis is to **employ a hybrid machine learning algorithm, in particular neural networks, to detect anomalies and attacks in CPS considering its physical model**.

However this is not the first time such a fusion is examined. In the literature various approaches of the fusion of neural networks with theory-based models were presented. Due to research, five different methods were found. First, an approach that add physic based features like [1]. Then, another approach generating huge datasets artificially and learn the machine learning model to find the suspicious values of sensors rather than deciding in which status is the system, like in [2]. Third, a method that reduces the dimensionality of the problem, like in [3]. Fourth, a method that works on data defined over graphs, like in [4], and finally an approach that decomposes the problem into smaller ones, like in [5].

The first mentioned approach takes into consideration a physical model of the system. Based on the sensors' data, the physical model will generate additional physical features. Finally, those physical features, along with sensor's data are passed to the machine learning algorithm. This algorithm was visualized in figure 0-1.

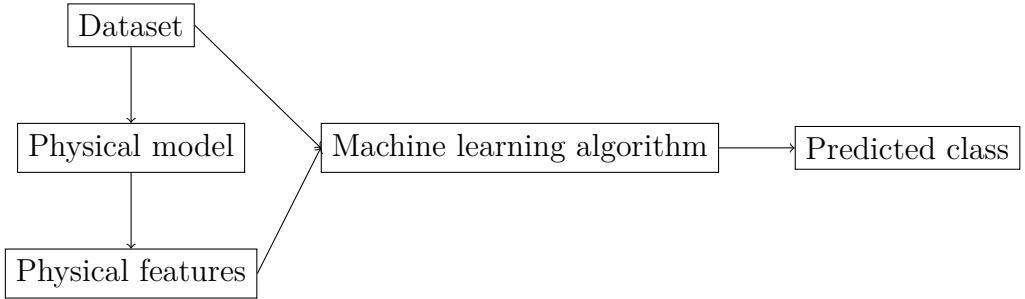


Figure 0-1: Physical features addition algorithm visualization

In [1] this approach was used to model water temperature in a lake at varying times and depths. As physical model, they used the state-of-the-art general lake model [6] to obtain the temperatures obtained via simulation. Moreover, their approach incorporates a loss-function composed of three main components: the empirical error, the structural error and the physical inconsistency. The loss-function is a function that the machine learning algorithm tends to minimize during the training process. The empirical error evaluates the difference between the predicted and expected values, the structural error is proportional to the model complexity, whereas the physical inconsistency evaluates if the predicted value is physically correct. The authors of [1] have proven that their approach gives better results than other attempts to resolve the same problem.

The second approach, on other hand, presented in [2], takes into consideration an attack generation engine. This engine takes as an input a configuration file describing the system then it creates thousands of samples with various features describing sensors readings and others physic based. Then, the attacks are simulated by spoofing the values of samples. The dataset obtained this way is then trained to a machine learning algorithm, which will be able to tell to which extent the values of the features are physically correct. A simplified illustration of the training process of the algorithm is presented in figure 0-2.

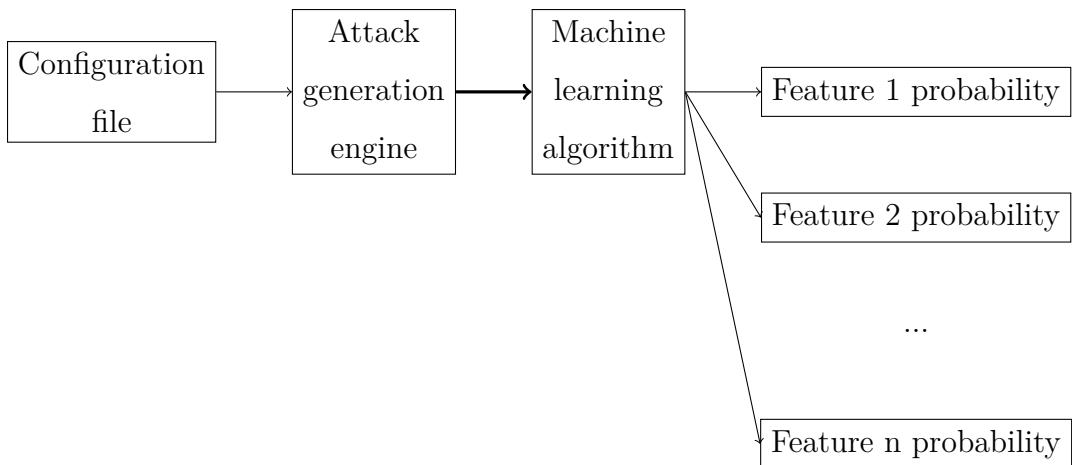


Figure 0-2: Attack generation approach training algorithm visualization

The authors of [2] presented their work on the example of a power system, in particular the IEEE

30-bus test case model. They achieved an accuracy of 99% with a neural network with a 60-node layer, which is a very good result.

The third approach, proposed by [3], uses so called proper orthogonal decomposition (POD) in order to reduce the dimensionality of the parameters of the system. This parametrization is then combined with machine learning algorithms to learn the relationship between the input and the data obtained from the POD. The details of the POD algorithm will not be discussed due to their complexity.

The authors of this last paper, used various machine learning techniques like neural networks or decision tree and they presented multiple applications of their approach, however all of them were in the area of fluids. The obtained results in each case showed a significant improvement of accuracy.

The fourth approach, proposed by [4] was described in particular on distribution system state estimation problem. Its role is to estimate the state of the system based on physical measurements. The accuracy of those measurements may vary, that is why they can be divided into smaller problems. This division into partitions is assured by using graphs to represent spatially the data. The state estimation at a certain bus of the network can be limited to using the measurements from bus' partition(s). That is why the main idea behind the used neural network is to set the weights of the measurements outside this partition(s) to zero. The authors found that their approach is more performant than Gauss-Newton approach, which rely only on physics.

The final approach, presented in [5], suggests to create modular neural networks to solve complex problems by dividing them into smaller ones. This way multiple neural networks are created and trained using only data concerning the small problem that they solve. The authors of [5] proposed in the same work many other approaches but they were already discussed.

All the discussed methods have one in common: they can only detect the malfunctioning in general, without differentiating between attacks and a normal fault. In this master thesis, an attempt will be made to develop an algorithm able to detect the system status - normal behaviour, a fault and an attack. In order to do that, the **scope of this thesis** includes a description of an example of cyber physical system, the description of different machine learning algorithms along with the comparison of their performances, the analysis of the importance of features and finally the proposal of three different solutions to the initial problem.

This thesis is composed of four chapters. The first chapter describes in details an example of a cyber-physical system provided with a ready to use datasets. The second chapter is a description of used machine learning toolkits, machine learning techniques and metrics for their evaluation. This chapter also includes a comparison of the values of those metrics in relation to the used techniques and toolkits. The third chapter is a description of different tools for the interpretation of results and

the proposal of a simple algorithm to extract the most important features for the false predictions made by a classifier. The last chapter presents a proposal of three different techniques to answer the goal of this master thesis. This chapter comes with an analysis of the results obtained after applying those techniques.

Chapter 1

Power system as a CPS example

In order to focus on the implementation of the hybrid machine learning algorithm, a CPS, with ready to use datasets, was chosen from a list provided in [7]: the **power system** [8], which network diagram was represented in figure 1-1. The system is composed of two power generators who are alimenting the whole system. Intelligent Electronic Devices (IEDs) R1 to R4 and the breakers BR1 to BR4 can be found connected directly to those generators. Each IED switches its corresponding breaker when a fault is detected, valid or fake. The communication between the IEDs and the Substation Switch is done wirelessly. On the other hand the Substation Switch is connected with the Primary Domain Controller (PDC) and the Control Room.

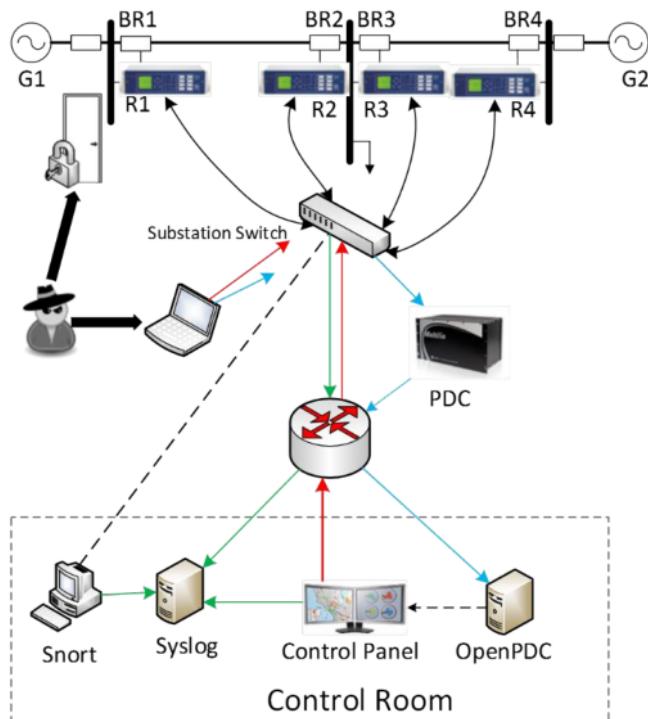


Figure 1-1: Power system network diagram [8]

The operation of this power system can be described following 6 main scenarios:

- normal behaviour,
- short-circuit,
- line maintenance,
- remotely opening the breakers (attack),
- disruption of fault protection system (attack),
- fault imitation (attack).

Each of those scenarios can be divided into several sub-scenarios concerning different entities of the system or/and the failure range. Every scenario was labelled with a number between 1 and 41. In this way **37 scenarios** are obtained, divided and numbered as follows:

- 1 no events scenario, its number is 41,
- 8 natural fault scenarios, its number ranges are 1-6 (short-circuit) and 13-14 (line maintenance),
- 28 attack scenarios, its number ranges are 7-12 (fault imitation), 15-20 (remotely opening the breakers), 20-30 and 35-40 (disruption of fault protection system).

The reason for dropping the numbers between 31 and 34 in the naming process of scenarios is not known.

The datasets provided in [7] represent **78377 events**, in which one of those scenarios was reproduced in the system. They have been grouped by scenario into 3 datasets: binary (attack or normal operation), three-class (attack, normal fault and no events) and multiclass (differentiating all 37 scenarios). Each of these 3 datasets is composed of 15 .arff or .csv files comporting in average 141 events for each of 37 scenarios. The exact number of events per file for each scheme is illustrated in figure 1-2. For the 3 class dataset **55663 attack**, **18309 natural fault** and **4405 normal operation** events were found. The distribution of these schemes throughout the files is shown in figure 1-3.

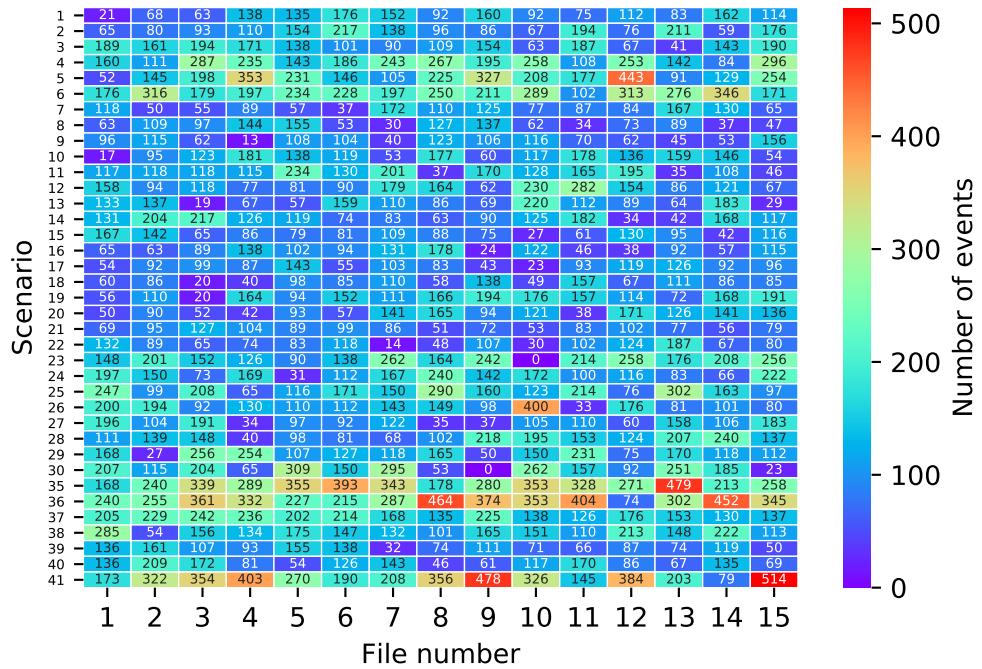


Figure 1-2: Scenarios distribution throughout all 15 files

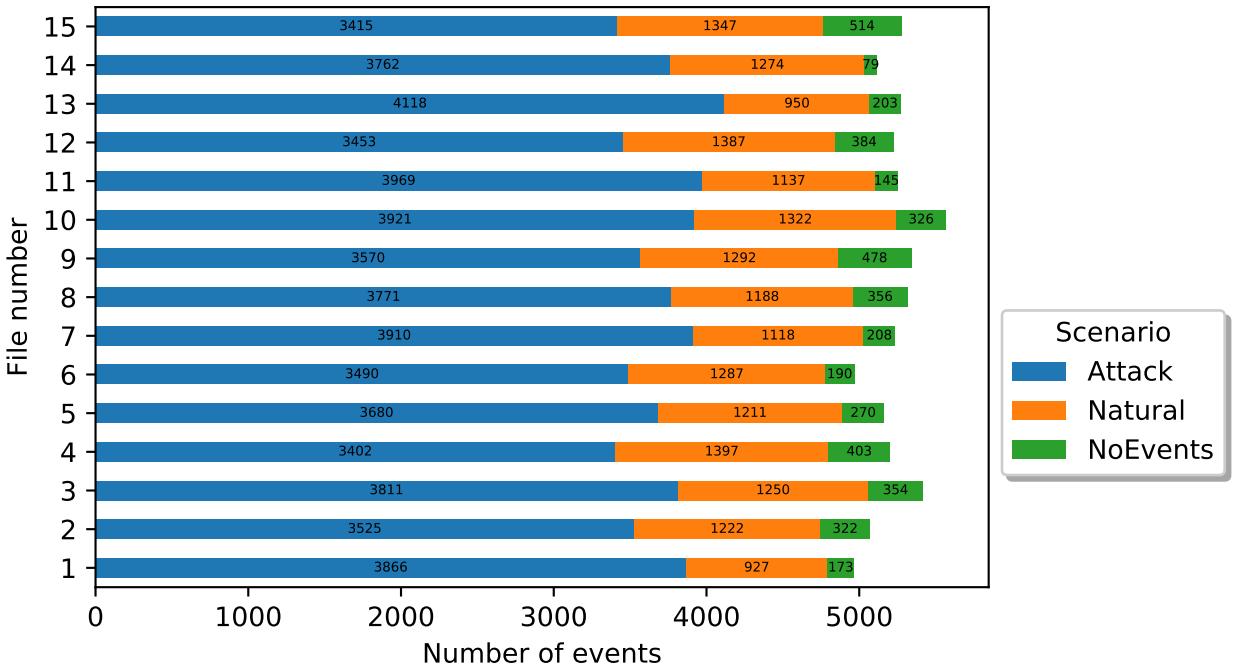


Figure 1-3: Scenarios distribution throughout the 3-class dataset files

Figure 1-3 shows also this distribution for the binary datasets. It is sufficient to add the number of natural (orange) and normal operation (green) events.

The scenarios are not equally distributed in the case of the 37 schemes dataset, it is especially shown by the standard deviation of 61, which is an important value. This is also visible because

some scenarios counting less than 100 events and others even 500. On the other hand, in the case of 3-class scenarios, the distribution is even more unequal compared to the 37 schemes dataset. The **mean standard deviation among all files is equal to 1767**, which is an enormous result.

Every electrical grid around the world uses a **3-phased** electric power. Such a grid is composed of three alternating current generators combined. Those generators pass the current in three conductors. Each of these conductors conducts a phase of current named A, B and C respectively. The current phases have the same frequency, but a difference of phase of $1/3$ of a cycle between each of them. In addition to that, each current has a corresponding voltage, with the same frequency and phases differences [9].

In order to simplify the analysis of three-phase power systems, symmetrical components transformation is used, for both voltage and current. This transformation is defined as:

$$\begin{bmatrix} V_0 \\ V_1 \\ V_2 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 \\ 1 & a & a^2 \\ 1 & a^2 & a \end{bmatrix} \begin{bmatrix} V_A \\ V_B \\ V_C \end{bmatrix}, \quad (1.1)$$

where \mathbf{V}_0 , \mathbf{V}_1 , \mathbf{V}_2 are called respectively **zero sequence**, **positive sequence** and **negative sequence**, $a = e^{j\frac{2\pi}{3}}$ and V_A, V_B, V_C the A-C voltage phases [10]. As each sequence is a weighted sum of sinusoidal functions (A-C phases), it can be on its own written as one sinusoidal function after mathematical transformations.

Each phase is a **sinusoidal** function. Its equation form is $y = A \cdot \sin(\omega t + \theta)$, where A is the amplitude, ω the angular frequency and θ the initial phase. Two terms will be used in what follows: the **magnitude** which is the absolute value of the amplitude and the **angle** which refers to initial phase. These two variables were illustrated in figure 1-4.

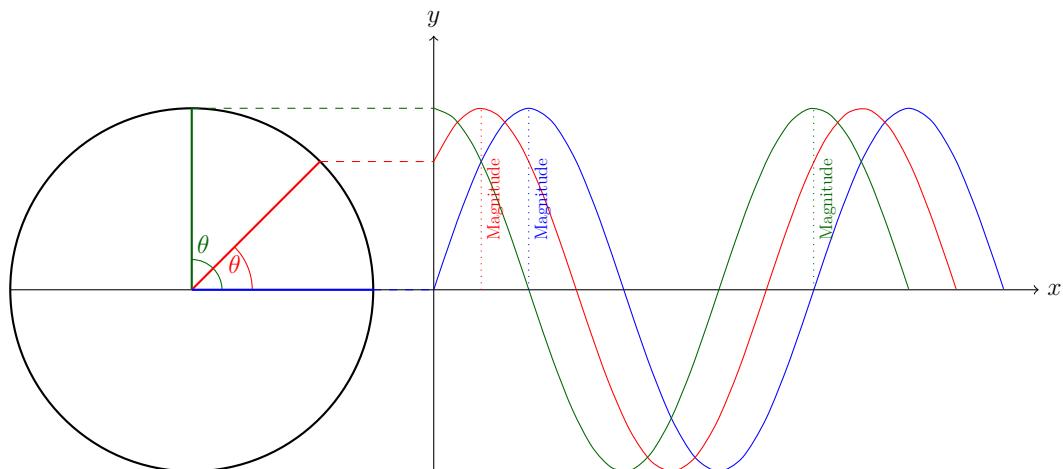


Figure 1-4: Magnitude and angle (θ) of sinusoidal functions

Moreover, when talking about electrical components, the term of impedance can come across. The impedance can be seen as a generalization of resistance of an electrical component. It is defined as the ratio of the voltage over current that passes through this component and takes a complex value of form $z = x + yi$. Every complex number can be also written in form of $z = e^{i\phi}$, where ϕ is the angle of the complex number z , and in this particular case the angle of the impedance.

Every previously mentioned event is described by **128 features**: 116 provided by four IEDs (each one provides 29 types of measurements) and 12 other features are reserved for control panel logs, snort¹ alerts, relay logs of 4 IEDs. The mentioned 116 features, each has a label formed by **concatenation of the source IED reference** (it can be R1, R2, R3, R4) and the **measurement name**, as provided in table 1.1. For example R4-PM5:I stands for phase B current phase magnitude measured by R4.

Table 1.1: IED measurements description [8]

Feature	Description
PA1:VH – PA3:VH	Phase A-C Voltage Phase Angle
PM1:V – PM3:V	Phase A-C Voltage Phase Magnitude
PA4:IH – PA6:IH	Phase A-C Current Phase Angle
PM4:I – PM6:I	Phase A-C Current Phase Magnitude
PA7:VH – PA9:VH	Pos.–Neg.–Zero Voltage Sequence Angle
PM7:V – PM9:V	Pos.–Neg.–Zero Voltage Sequence Magnitude
PA10:VH - PA12:VH	Pos.–Neg.–Zero Current Sequence Angle
PM10:V - PM12:V	Pos.–Neg.–Zero Current Sequence Magnitude
F	Frequency for relays
DF	Frequency Delta (dF/dt) for relays
PA:Z	Appearance Impedance for relays
PA:ZH	Appearance Impedance Angle for relays
S	Status Flag for relays

Those datasets have been used in several works related to CPS cyber-attack classification, one of which is [11], where the author tries to find the most accurate algorithm to predict the status of the power system.

¹Snort - Network Intrusion Detection and Prevention System

In this chapter the power system was presented as an example of a cyber-physical system. The datasets including the sample data of this system were presented and described in details, including the distribution of the classes along samples and the exact physical explanation of all the features that compose those samples. This system is used in the rest of this master thesis as the studied case. The next chapter will concern the evaluation of different machine learning techniques regarding their capability to correctly predict the statuses of this CPS.

Chapter 2

Machine learning algorithms comparison

Before going further and analysing neural networks in order to create a hybrid machine learning algorithm for anomaly detection in the power distribution system presented in chapter , a deeper look at classical machine learning algorithms will be taken, in particular Random Forest, Support Vector Machine (SVM), NaïveBayes and multilayer perceptron. However this was done before in [11] using the black-box model algorithms only. In their approach they used Weka [12] in order to find the most performant algorithm among 7 they have chosen (OneR, NNge, Random Forest, Naïve Bayes, SVM, JRipper, Adaboost).

This chapter shows an attempt to reproduce the results provided in [11], using two different machine learning toolkits (Weka and scikit-learn [13]) in order to confirm the obtained results. That is why, first, these two toolkits will be presented, then the obtained results will be discussed.

2.1 Machine learning toolkits

2.1.1 Weka

Weka, or more exactly **Waikato Environment for Knowledge Analysis**, is an open source machine learning software developed at The University of Waikato in Hamilton, New Zealand and based on Java programming language. It is well known especially in academic environments and a lot of machine learning researches were conducted using it, one of them the mentioned before [11]. It incorporates various machine learning tools: classifiers, regressors, visualizers, data pre-processor etc...

Weka is characterised by 3 main operating schemes. First, it can be run using a **graphical user interface (GUI)**, enabling the user, even without deep knowledge in programming, to make machine learning experiments and analyse available classifiers. Second, more advanced users have the option

to run all the available tools using a **command line**. Finally, Weka's tools can be **integrated directly into code in several programming languages (Java, Python, R, Spark)**, which enables even larger versatility.

2.1.2 scikit-learn

scikit-learn is an open-source machine learning **Python library** developed originally by David Cournapeau as a Google Summer of Code project and now it is maintained by a team of volunteers. It is well known in both academic and commercial environments, since it is used by many enterprises such as Spotify, Evernote, Booking.com, and research facilities like Inria or Télécom ParisTech [14]. It includes, just like Weka, various machine learning tools such as classifiers, regressors, data pre-processor etc... Its visualisation capabilities are limited, however there exist many additional Python packages for data visualisation such as YellowBrick, Eli5 and others... They will be further discussed in next chapter.

scikit-learn can be used only as an extension of Python language, what makes a bit harder for non experts to start working with it. However, since Python is an user friendly programming language and scikit-learn has a well made documentation, this toolkit is easy to use. Along with other Python packages, and especially pandas, scikit-learn is a very powerful, ergonomic and versatile solution for machine learning problems.

It might be also interesting to mention that scikit-learn can be used within Weka after installing the appropriate add-on. This enables Weka users to use scikit-learn classifiers and regressors and run Python code just inside Weka's GUI.

2.2 Used machine learning algorithms

Before talking about the comparison of machine learning algorithms, the used classifiers are briefly described in this section in order to better understand how do they work. In parallel, values of classifiers' parameters used during all the tests are presented in tables 2.1, 2.2 and 2.3, and, in majority, they represent the default values found in Weka, in order to reproduce the results from [11].

2.2.1 Random Forest

Random Forest, as the name indicates, is an **ensemble** of a given number of trees, and more exactly **decision trees**. This given number of trees is set using the parameter *numIterations* in Weka and *n_estimators* in scikit-learn. Each of decision trees is created based on a randomly chosen set of

features of the dataset. The number of features is given by the parameter *numFeatures* in Weka and *max_features* in scikit-learn, which indicate the way it is calculated - in this case as the logarithm of base 2 of the number of features.

On the other hand, decision trees represent a structure capable of determining the class of the predicted sample. It is composed of nodes connected to each other in a form of tree. Each node corresponds to a condition related to the features, for example if a particular feature is higher than a certain number. Each node, except the decision ones, has two child nodes corresponding to the answer as yes or no to the condition from parent node. For each of two answers, particular classes are assigned, for example, given the case study in this work, yes answer can correspond to classes Attack and Natural and the answer no to class NoEvents. When running a tree on a particular sample, the nodes are followed given the answers on conditions, until arriving to the decision node. The class corresponding to this node is taken as the final output of the decision tree. The number of layers of nodes is called the depth of the tree and it is set using the parameter *maxDepth* in both Weka and scikit-learn.

The choice of condition in a node is based on calculating **gini impurity** or **information gain** metric for each feature and determining, based on one of them, which feature has the biggest effect on the choice of the class. Gini impurity determines the probability of a feature being classified incorrectly when selected randomly, it is given by the equation:

$$\text{Gini impurity} = 1 - \sum_{i=1}^n (p_i)^2, \quad (2.1)$$

where n is the number of classes and p_i the probability of the feature being classified for the class i . The information gain on the other hand, determines the quantity of information that the features gives about a particular class, it is calculated as:

$$E(S) = \sum_{i=1}^n -p_i \log_2 p_i. \quad (2.2)$$

Weka uses the information gain metric, while in scikit-learn it is possible to choose the metric using the parameter *criterion*, which by default is set to gini impurity, which is less computationally complex.

The random forest, composed of a number of those decision trees, determines the class of a sample taking the class chosen by the biggest amount of decision trees. In both Weka and scikit-learn, this classifier comes with more, less crucial parameters, which are all listed in table 2.1.

Table 2.1: Random Forest classifier used parameters in Weka and scikit-learn

(a) Weka [15]		(b) scikit-learn [16]	
Parameter	Value	Parameter	Value
bagSizePercent	100	n_estimators	100
batchSize	100	criterion	"gini"
breakTiesRandomly	False	max_depth	None
calcOutOfBag	False	min_samples_split	2
computeAttributeImportance	False	min_samples_leaf	1
debug	False	min_weight_fraction_leaf	0.0
doNotCheckCapabilities	False	max_features	"log2"
maxDepth	0	max_leaf_nodes	None
numDecimalPlaces	2	min_impurity_decrease	0.0
numExecutionSlots	1	min_impurity_split	None
numFeatures	0	bootstrap	True
numIterations	100	oob_score	False
outputOutOfBagComplexityStatistics	False	n_jobs	None
printClassifiers	False	random_state	None
seed	1	verbose	0
storeOutOfBagPredictions	False	warm_start	False
		class_weight	None
		ccp_alpha	0.0
		max_samples	None

2.2.2 SVM

SVM, or more exactly **Support Vector Machine**, is a classifier that creates an N-dimensional space, in which all the samples from the training dataset are put, and then divided, geometrically, into regions, where each region corresponds to one class. The N dimensions of this space corresponds to all the features of the datasets and their transformations (for example a square of one of the features). Those transformations are done using a kernel function which can be set in Weka and scikit-learn using the *kernelType/kernel* parameter. Three kernel types are more commonly used: linear, polynomial and radial basis function (exponent based). The default kernel in both Weka and scikit-learn is radial basis function. Each type of kernel function comes with a set of parameters itself, like the degree in the case of polynomial and all of them can be set through the parameters of both used machine learning toolkits.

In fact, the transformations do not calculate additional dimensions for every sample, but calculate the distance between points in the N-dimensional space. This distance is mathematically defined

as the **dot product** of the two points. The dot product on other hand is defined of the sum of products of coordinates for every dimension.

The division is made by determining so called **hyperplane**, which for example in a 2D features space is just a line. The points from each class that are the closest to the hyperplane are called **support vector points**, from where comes the name of this classifier [17] [18].

The algorithm is running within a set number of iterations (can be set within Weka and scikit-learn) dividing the training dataset to smaller tranches using the cross-validation technique. The goal is to find the hyperplane that way so the number of misclassified samples is the smallest in that particular iteration.

SVM in Weka and scikit-learn comes additionally with other parameters than are listed on table 2.2.

Table 2.2: SVM classifier used parameters in Weka and scikit-learn

(a) Weka [19]		(b) scikit-learn [20]	
Parameter	Value	Parameter	Value
SVMType	C-SVC (classification)	C	1.0
batchSize	100	kernel	"rbf"
cacheSize	40.0	degree	3
coef0	0.0	gamma	"scale"
cost	1.0	coef0	0.0
debug	False	shrinking	True
degree	3	probability	True
doNotCheckCapabilities	False	tol	1e-3
doNotReplaceMissingValues	False	cache_size	7000
eps	0.001	class_weight	None
gamma	0.0	verbose	False
kernelType	radial basis function	max_iter	1000
loss	0.1	decision_function_shape	"ovr"
modelFile	Weka-3-8-4	break_ties	False
normalize	False	random_state	None
nu	0.5		
numDecimalPlaces	2		
probabilityEstimates	True		
seed	1		
shrinking	True		
weights			

2.2.3 NaïveBayes

NaïveBayes is a classifier based on the **Bayes theorem**, which states that the probability of event A, given that the event B happened is equal to the product of the probability of B given that A occurred and the probability of A, divided by the probability of B:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \quad (2.3)$$

While the naivety of the algorithm is due to it does not taking into consideration the eventual relations between the features.

The NaïveBayes classifier can be both **multinomial** and **gaussian**. In the multinomial case the probabilities are calculated the classical way (number of occurrences over all samples), whereas in the gaussian case, the probabilities are taken from the gaussian distribution of a given event. For both Weka and scikit-learn the gaussian case were used.

The list of all the parameters available is listed on table 2.3, however they impact on how the algorithm works is marginal.

Table 2.3: NaïveBayes classifier used parameters in Weka and scikit-learn

(a) weka [21]		(b) scikit-learn [22]	
Parameter	Value	Parameter	Value
batchSize	100	priors	None
debug	False	var_smoothing	1e-9
displayModelInOldFormat	False		
doNotCheckCapabilities	False		
numDecimalPlaces	2		
useKernelEstimator	False		
useSupervisedDiscretization	False		

2.2.4 Multilayer perceptron

The multilayer perceptron is a classifier inspired by the structure of the human brain. It is constituted from a set of interconnected neurons, which are divided into at least 2 layers: input layer, n hidden layers and output layer. Each layer consist of a number of neurons. The input layer has always a number of neurons equal to the number of features, and the output layer has always this number equal to the number of outputs of the problem. That is why, in the analysed case the input layer has 128 neurons and the output layer only one neuron corresponding to the status of the power plant (normal behaviour, natural fault, attack). Each of the mentioned neurons has a weight. The

neurons are interconnected. The output of a neuron is considered as the input for the neurons of the following layer. Those neurons of the following layer calculates the weighted sum of the inputs and passes the results through a function, called activation function. The result obtained this way is considered as the output of the given neuron and goes to the following layer. The only exception is the input layer, which input is the values of features and its neurons just copy the input to the output [23] [24] [25].

The training process consists in updating the weights of the neurons during a predefined number of iterations called epochs. Before starting the training process the initial weights w_i values for hidden and output layers should be chosen. This choice can be done empirically or randomly. Moreover, a real positive number η called training step must be fixed. In each epoch and for each training pattern, the weights are updated following the equation:

$$w_i = w_i + \eta \cdot (z^{(\mu)} - y) \cdot x_i^{(\mu)}, \quad (2.4)$$

where $z^{(\mu)}$ is the desired output for the μ -th training sample, y is the neuron output (weighted sum passed through the activation function), $x_i^{(\mu)}$ is the neuron's input value for the same μ -th training sample and i indicates the neuron index in the trained layer. After each epoch, the training samples are shuffled [25].

The scikit-learn implementation of multilayer perceptron classifier is limited because it does not give a full control of each hidden layer and its activation function. It integrates however the possibility to determine the number of epochs called iterations and the training step called learning rate. All available parameters available are shown in table 2.4. The implementation of MLP exists also in Weka, but it is as well limited, for this reason the analysis of its results will not be taken into consideration.

There exists multiple python frameworks offering much more flexibility when creating neural networks like PyTorch [26] or Keras framework [27], which were created in order to offer the possibility to create complex neural networks able to be used in large-scale applications. Even the developers of scikit-learn state on their website [24]:

This implementation is not intended for large-scale applications. In particular, scikit-learn offers no GPU support.

The mentioned frameworks will not be used in this work because this flexibility is not required in this case. However, it does not mean that it is not possible to make the results even better using those frameworks.

Table 2.4: MLP classifier parameters in scikit-learn [28]

Parameter	Value
hidden_layer_sizes	(20,)
activation	'relu'
solver	'adam'
alpha	0.0001
batch_size	'auto'
learning_rate	'constant'
learning_rate_init	0.001
power_t	0.5
max_iter	1000
shuffle	True
random_state	None
tol	$1e - 4$
verbose	False
warm_start	False
momentum	0.9
nesterovs_momentum	True
early_stopping	True
validation_fraction	0.1
beta_1	0.9
beta_2	0.999
epsilon	$1e - 8$
n_iter_no_change	10
max_fun	15000

2.3 Metrics for classifiers comparison

All those classifiers give predictions that can be divided into 4 main groups. Assuming for a moment, for illustration purposes, that normal fault class is positive and attack class is negative, it could be differentiated between:

- true positive predictions (**tp**): correct prediction of the normal fault class,
- true negative predictions (**tn**): correct prediction of the attack class,
- false positive predictions (**fp**): incorrect prediction of normal fault class,
- false negative predictions (**fn**): incorrect prediction of attack class.

For that two classes, several metrics can be defined, that will be useful to compare between the used classifiers:

- **accuracy**: ratio of correct classifications over the total number of samples, or in other words:

$$\text{accuracy} = \frac{tp + tn}{tp + tn + fp + fn} \quad (2.5)$$

- **precision**: ratio of correct classifications for a particular class over all classifications that indicated that class, or:

$$\text{precision} = \frac{tp}{tp + fp} \quad (2.6)$$

- **recall**: ratio of correct classifications for a particular class, over all samples corresponding for this class, or:

$$\text{recall} = \frac{tp}{tp + fn} \quad (2.7)$$

- **f-measure**: weighted average of precision and recall given by the equation:

$$\text{f-measure} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.$$

Given that those metrics work only with binary problems with a positive and a negative class, for multiclass problems the **accuracy is always calculated as the ratio of true classifications over all classifications**, however for precision, recall and f-measure, the average is calculated, and that in 3 different ways [29]:

- **micro** average: the metrics are determined globally by calculating true positives as all correct predictions, false negatives and false positives as all incorrect predictions (assuming two classes A and B, when A is misclassified as B is a false positive for A, and in the same time a false negative for B). In this case precision, recall, f-measure and accuracy have exactly the same value,
- **macro** average: the metrics are calculated for each class, the concerned class is considered as positive, while the sum of others as negative. Then their arithmetic mean value is calculated,
- **weighted** average: the metrics are calculated for each class, then it calculates their weighted average value by the number of true instances for each class.

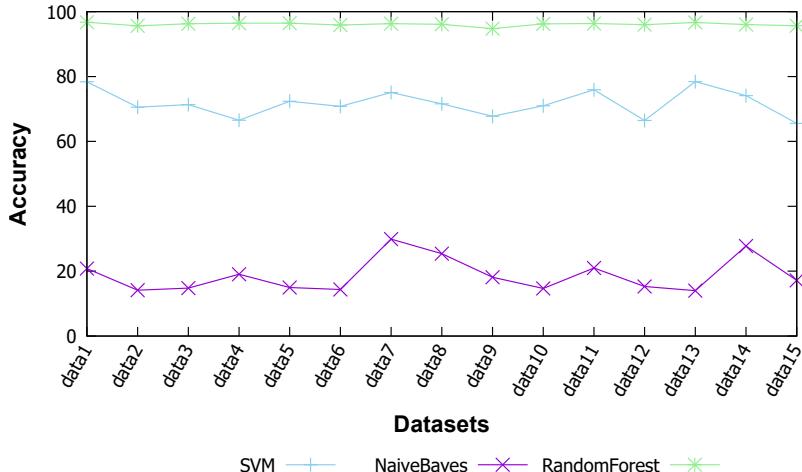
scikit-learn supports the calculation of precision, recall and f-measure in the 3 different ways explicitly, however in Weka only for f-measure all 3 are available, while for precision and recall there exist only weighted averages, in addition the metrics can be calculated for a particular class.

2.4 Comparison results

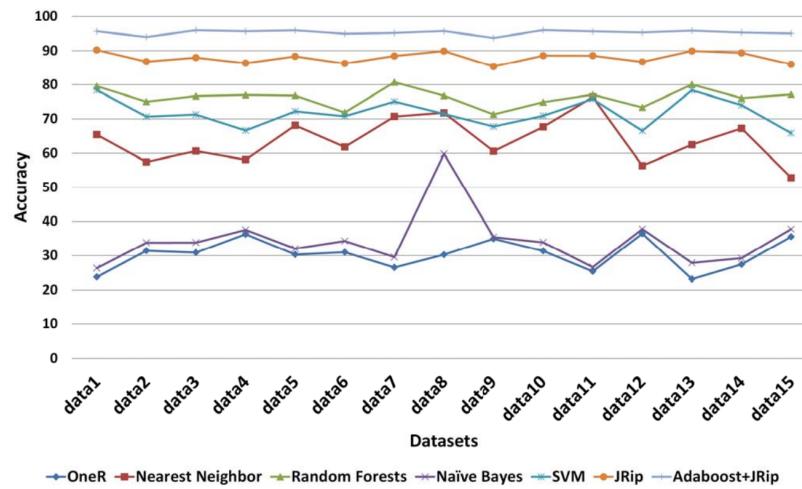
In order to reproduce the results presented in [11], Weka version 3.8.4 and scikit-learn version 0.23.1, running on Anaconda 3.18.11 with Python 3.7.6.final.0, were used. SVM in Weka comes from a package untitled *libsvm* available for this toolkit. The results are displayed in the same order as in [11], thus, the accuracy was presented in figures 2-1, 2-2 and 2-3 for the three class, binary and all 37 class cases respectively. Then, the precision, recall and f-measure can be found in figures 2-4, 2-5 and 2-6. For these three metrics the weighted average was used to obtain the results. For NaïveBayes the precision and f-measure could not be extracted using Weka for unknown reason.

Figure 2-1 shows the accuracy values per dataset file for the three-class case using Weka, scikit-learn and compared to the original results for different classifiers. An accuracy superior to 50% is observed for SVM and Random Forest classifiers using Weka and for MLP and Random Forest using scikit-learn. The original results show that all the classifiers have an average accuracy superior to 50% except Naïve Bayes and OneR classifiers. Figure 2-2 shows the accuracies in the binary case. The observations are similar, except for scikit-learn, where Naïve Bayes has this time an average accuracy superior to 50%. Figure 2-3 shows the accuracies for the multiclass datasets. The accuracies superior to 50% are Random Forest and SVM for Weka, Random Forest only for scikit-learn and Adaboost+JRip, JRip and Random Forest for the original results.

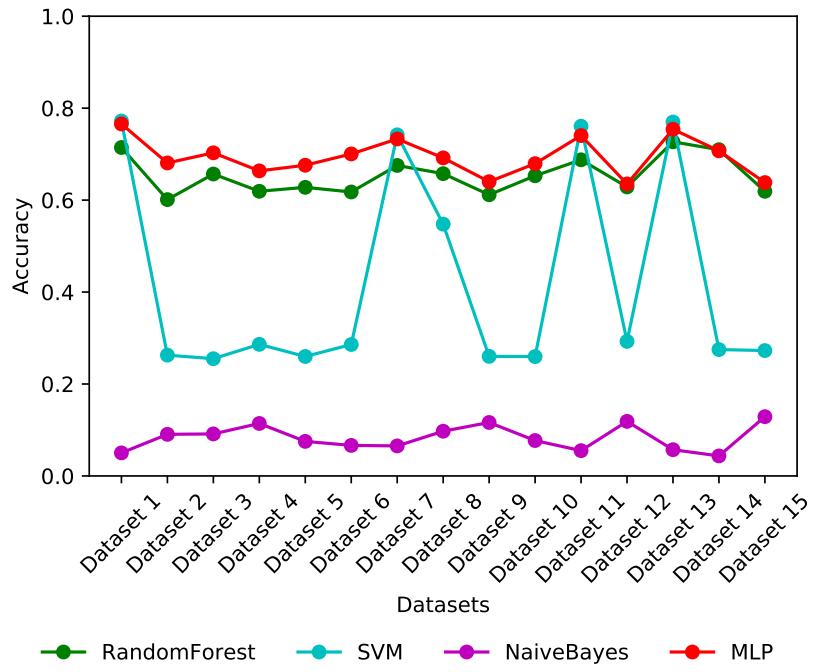
Figure 2-4 shows the precision metric values for different classifiers calculated using Weka and scikit-learn and compared to the original results. A precision superior to 0.5 is observed for all classifiers using Weka and and for Random Forest and partially for MLP (binary and three-class), SVM (binary and three-class) and Naïve Bayes (binary) using scikit-learn. The original results, on other hand, show a precision superior to 0.5 for Random Forest, JRIP and Adaboost+JRIP and additionally for the binary case using SVM. Figure 2-5 shows the recall metric values. A recall superior to 0.5 is observed for Random Forest and SVM in Weka’s results, for Random Forest, binary class Naïve Bayes, binary and 3-class MLP in scikit-learn’s results. The original results show a value superior to 0.5 for OneR, Naïve Bayes, JRip and Adaboost+JRip classifiers. Finally, figure 2-6 shows the f-measure values. The results are similar to those illustrating recall values for Weka and scikit-learn. On other hand, for the original results, only JRip and Adaboost+JRip classifiers show a value of f-measure superior to 0.5.



(a) Our attempt - Weka

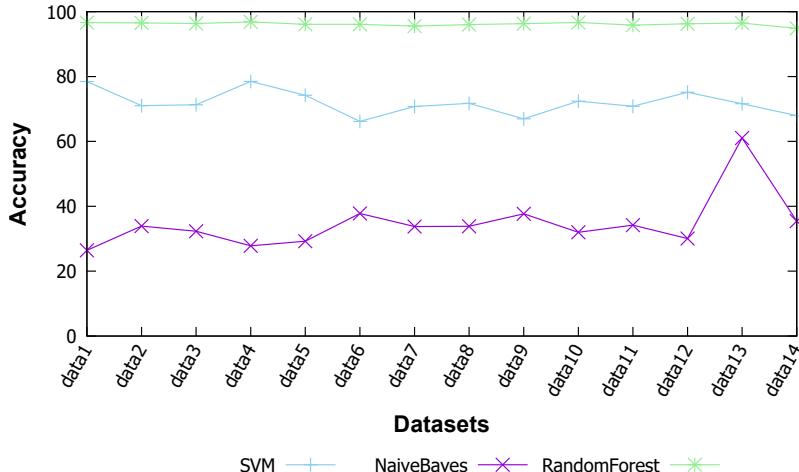


(b) Original results [11]

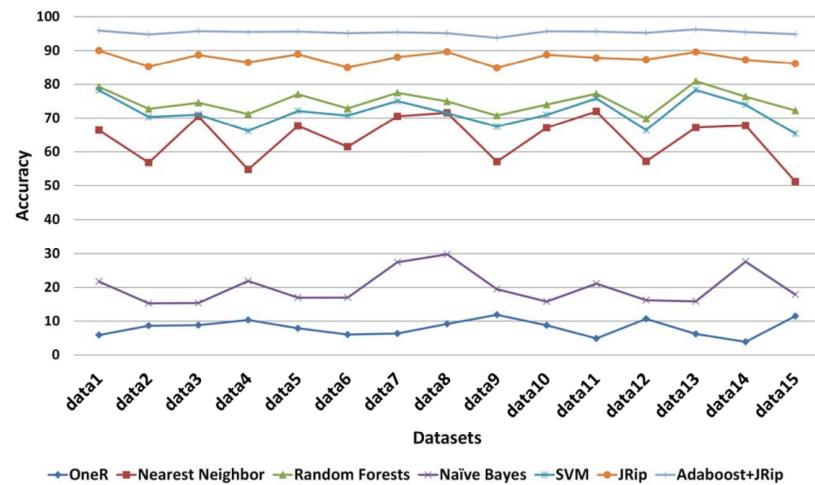


(c) Our attempt - scikit-learn

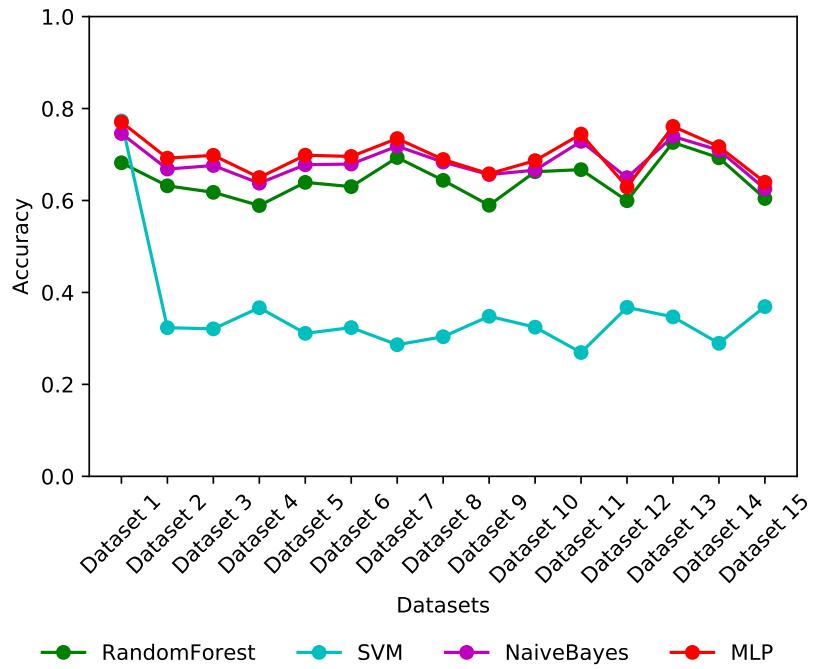
Figure 2-1: Plots showing the accuracy values per dataset file for different machine learning techniques using three-class datasets of the power system calculated using Weka and scikit-learn and compared to results from [11]



(a) Our attempt - Weka

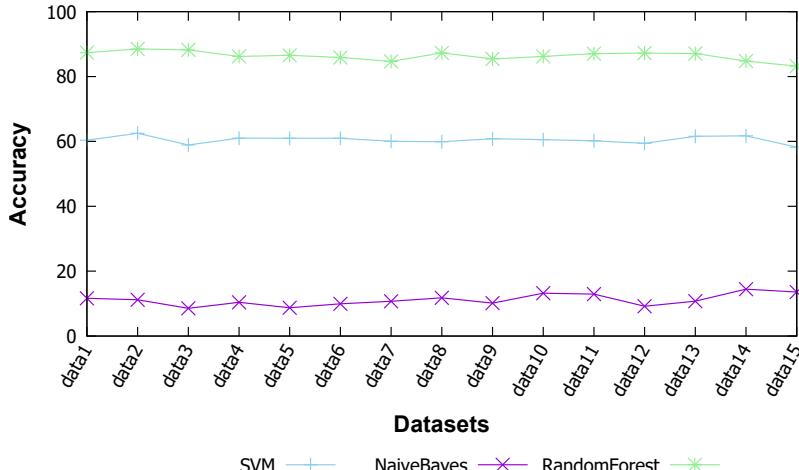


(b) Original results [11]

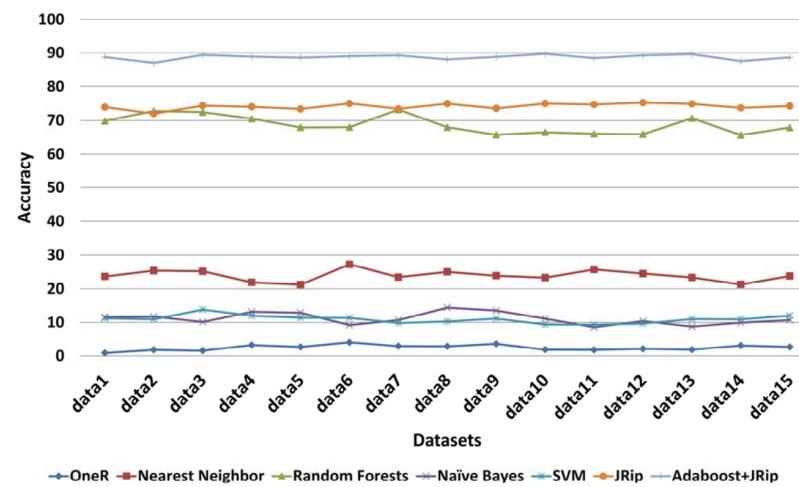


(c) Our attempt - scikit-learn

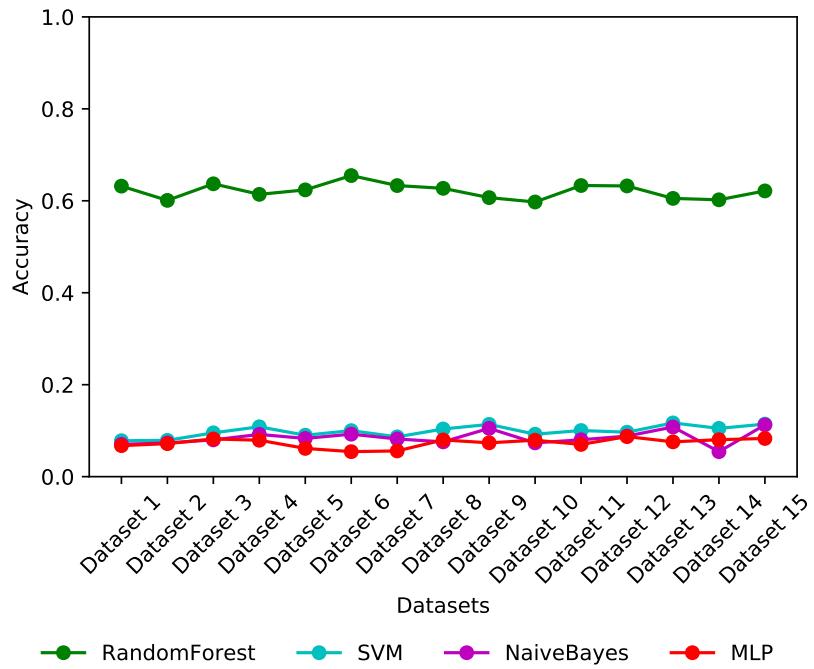
Figure 2-2: Plots showing the accuracy values per dataset file for different machine learning techniques using binary datasets of the power system calculated using Weka and scikit-learn and compared to results from [11]



(a) Our attempt - Weka

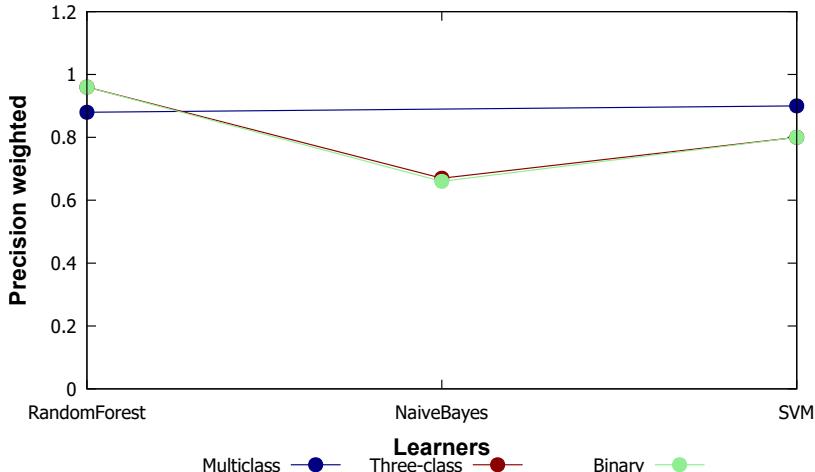


(b) Original results [11]

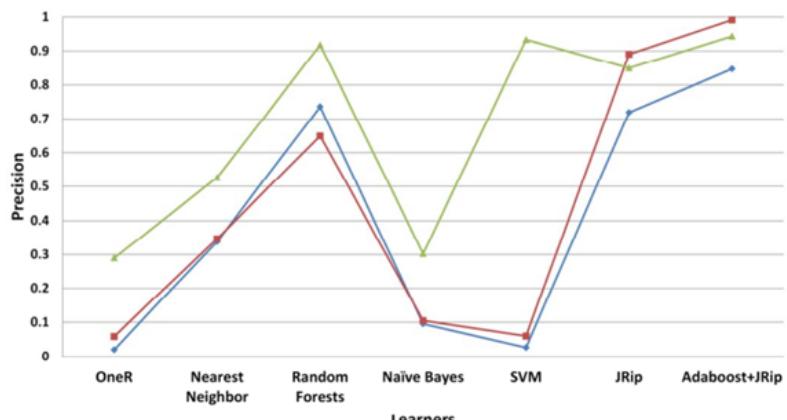


(c) Our attempt - scikit-learn

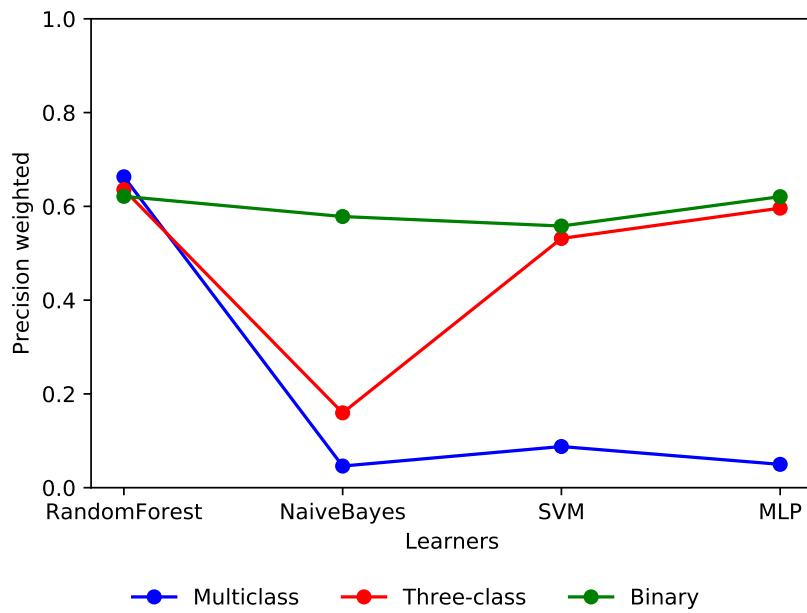
Figure 2-3: Plots showing the accuracy values per dataset file for different machine learning techniques using multiclass datasets of the power system calculated using Weka and scikit-learn and compared to results from [11]



(a) Our attempt - Weka

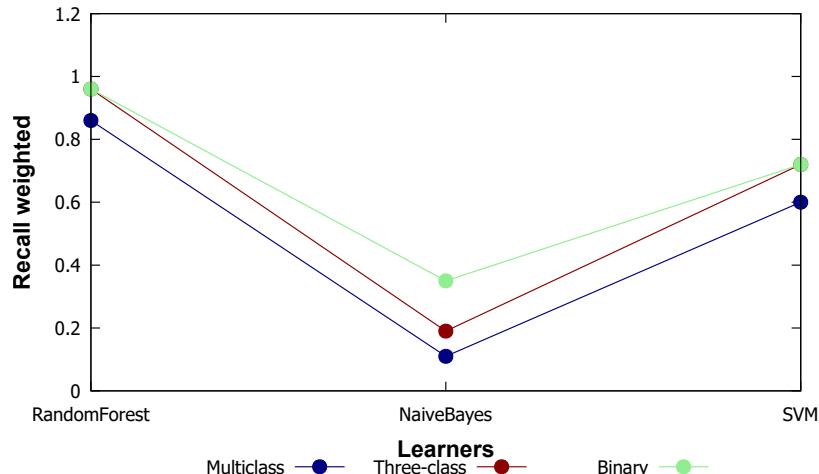


(b) Original results [11]

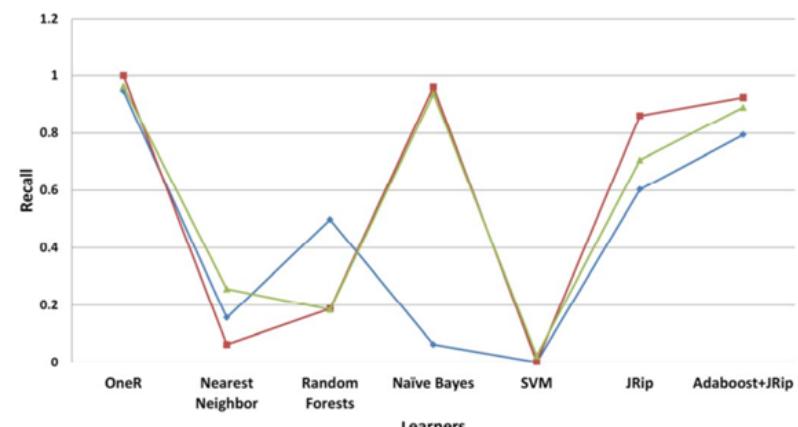


(c) Our attempt - scikit-learn

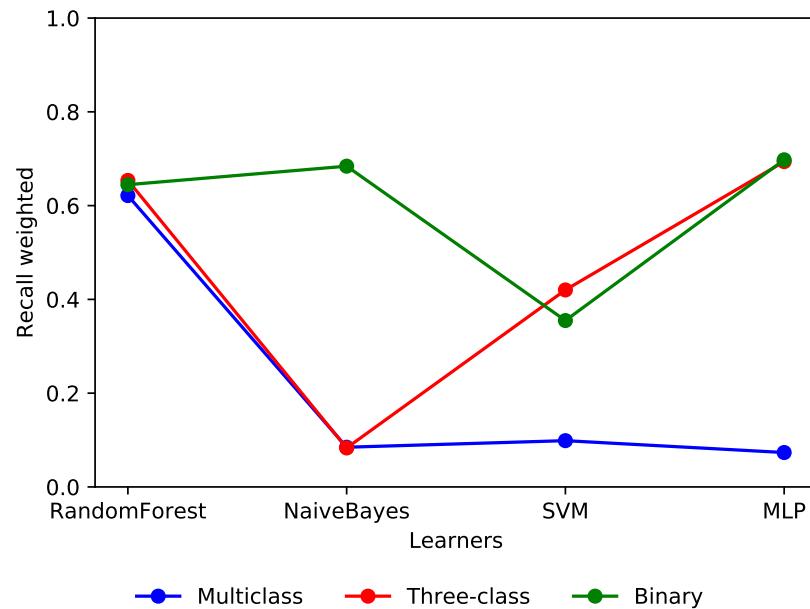
Figure 2-4: Plots showing the precision metric values for different machine learning techniques using different datasets of the power system calculated using Weka and scikit-learn and compared to results from [11]



(a) Our attempt - Weka

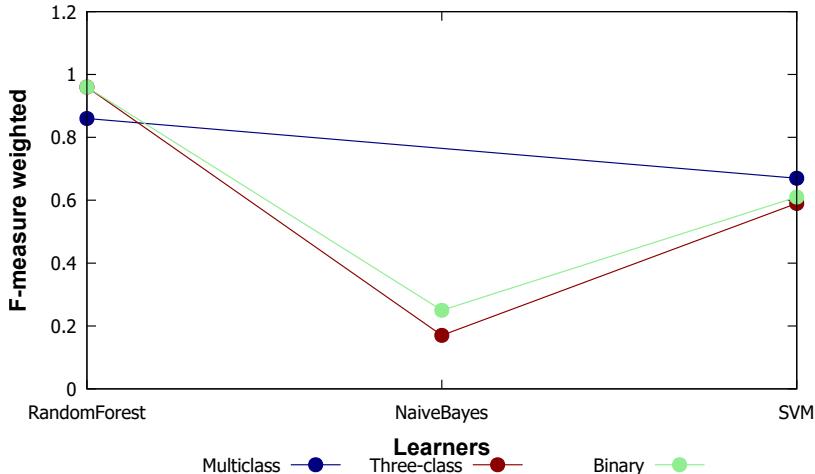


(b) Original results [11]

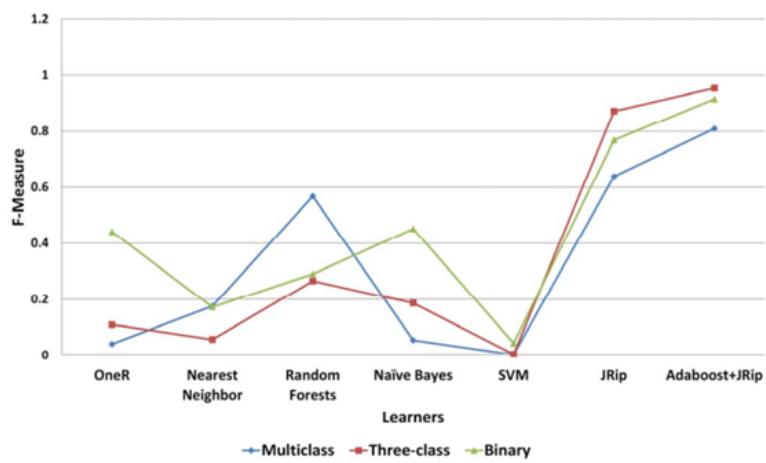


(c) Our attempt - scikit-learn

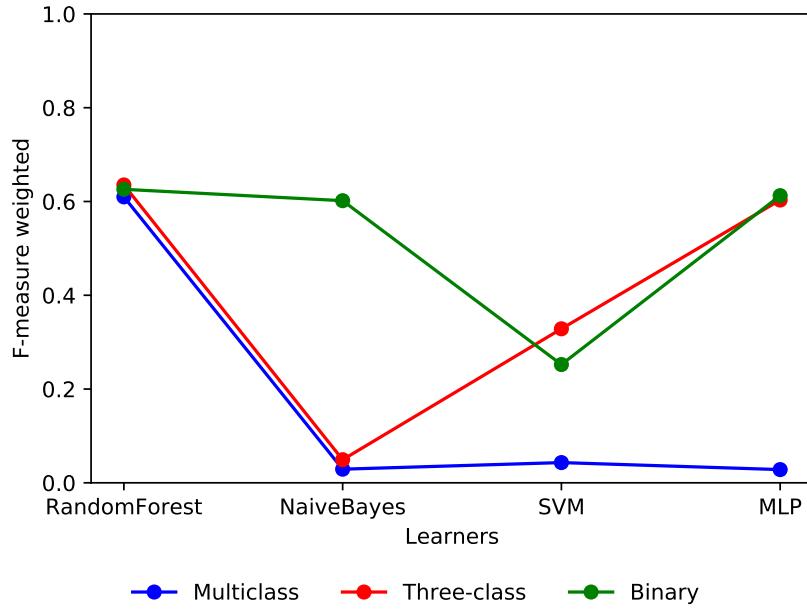
Figure 2-5: Plots showing the recall metric values for different machine learning techniques using different datasets of the power system calculated using Weka and scikit-learn and compared to results from [11]



(a) Our attempt - Weka



(b) Original results [11]



(c) Our attempt - scikit-learn

Figure 2-6: Plots showing the f-measure metric values for different machine learning techniques using different datasets of the power system calculated using Weka and scikit-learn and compared to results from [11]

The accuracies in the three-class case are quite similar on the three plots, except SVM in scikit-learn, what shows a huge discrepancy of accuracy between datasets. Moreover, a remarkably higher values of accuracy for Random Forest classifier are observed in the attempt using Weka, compared to other plots.

In the binary and multiclass cases, the Random Forest classifier in the attempts using Weka always shows higher values comparable to other plots. Scikit-learn, on other hand, for Random Forest classifier, gives results similar to those made by [11]. For SVM classifier, the discrepancy, this time, in the plots made by scikit-learn, is remarkably less important and the values oscillate around a constant value. For the binary case, this value is about 0.3 in scikit-learn, while in Weka the obtained value is 0.75, which makes it comparable to the original results in [11]. On the other hand, in the multiclass case, the value oscillates around 0.1, like it is the case in the results from [11]. In the attempt made in Weka the accuracy is much higher at about 0.6. For the NaïveBayes classifier the results are comparable, except the results obtained in scikit-learn in the binary case.

For precision, recall and f-measure, some similarities can be observed especially in the case of Random Forest classifier and in the case of NaïveBayes for multiclass. The results are generally different for the three metrics. For binary and multi-class cases the values of metrics are similar, except NaïveBayes classifier case. The most similar to the results from [11] are the results obtained using scikit-learn, despite the differences. The more efficient classifier is certainly Random Classifier, especially when using Weka. However, MLP classifier gives also quite good results.

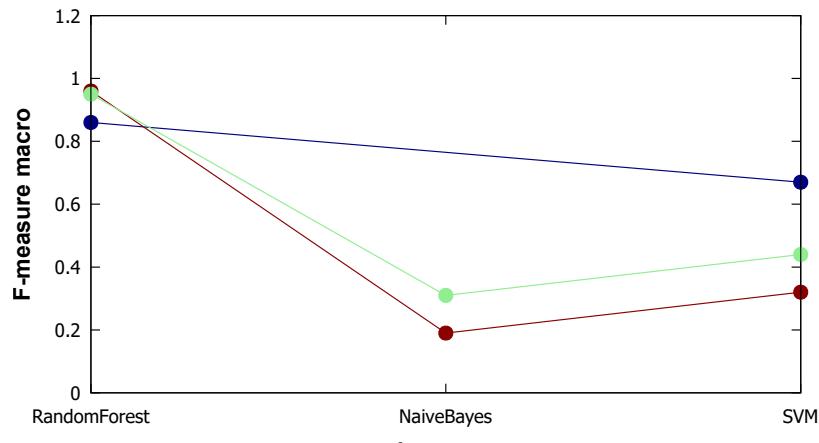
Moreover to get more information about the behaviour of the classifiers the f-measure macro and micro averages were calculated. The results are shown in figures 2-7, 2-8 respectively. These metrics were not presented in [11].

Figure 2-7 shows the f-measure macro average values for the three examined classifiers in Weka and four in scikit-learn. The plots show a value greater than 0.5 for Random Forest in the two toolkits, for SVM in multiclass case in Weka and for MLP in binary case in scikit-learn. Figure 2-8, on other hand, shows the f-measure micro average for those classifiers. Values superior than 0.5 are observed for Random Forest for both Weka and scikit-learn, for SVM in Weka, for Naïve Bayes in binary case using scikit-learn and for MLP in binary and three-class cases using scikit-learn.

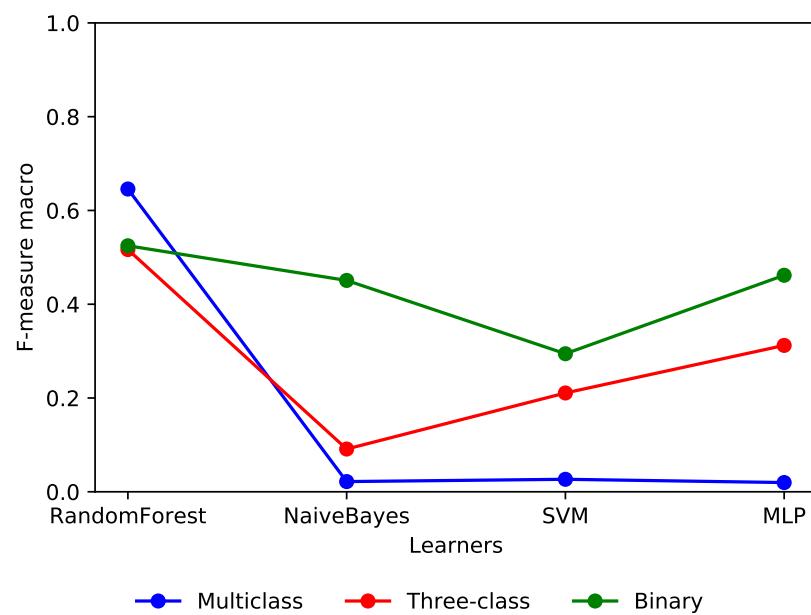
It would be hard to say that the obtained results in Weka and scikit-learn are similar. These results confirm that Weka and scikit-learn tend to produce different results and that results obtained by [11] are hard to reproduce because of lack information about the used parameters by the author. Additionally, it may be concluded that the best classifier is Random Forest, followed by MLP, especially in binary and three-class cases.

Since the results for 3-class and binary cases were comparable and the 3-class case gives more

information about the status of the system, in what follows, the focus will be on the 3-class case.

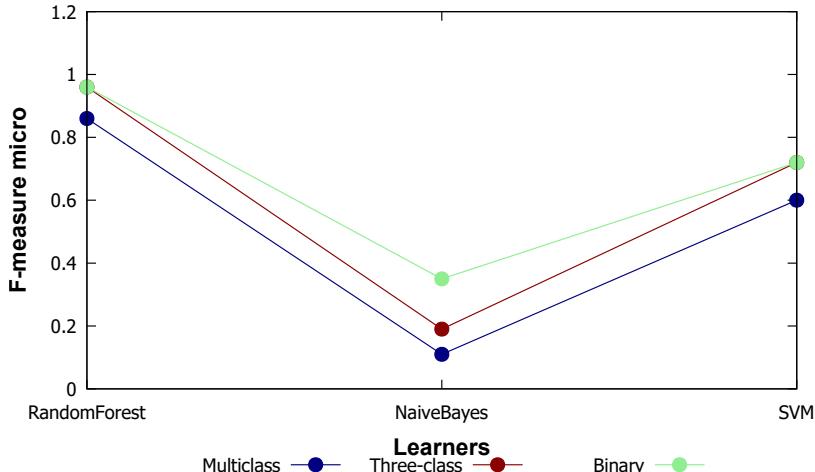


(a) Weka

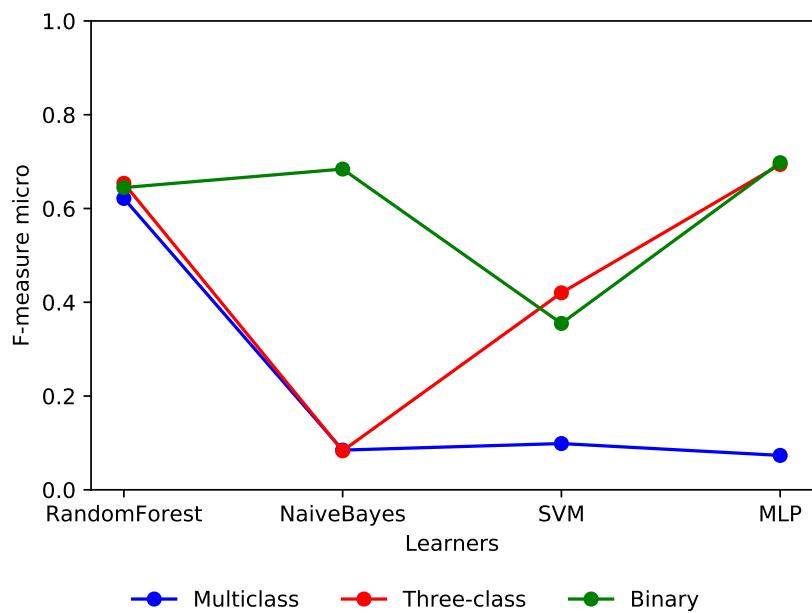


(b) scikit-learn

Figure 2-7: Plots illustrating F-measure macro average values for different classifiers using different datasets of the power system made using Weka and scikit-learn



(a) Weka



(b) scikit-learn

Figure 2-8: Plots illustrating F-measure micro average values for different classifiers using different datasets of the power system made using Weka and scikit-learn

2.5 scikit-learn further methods' analysis

In addition to all that, scikit-learn enables the user to plot the receiver operating characteristic (ROC) curves for each class and the confusion matrix. The ROC curve represents the plot of true positive rate when the false positive rate changes. It is based on the prediction probabilities for each class given for each tested sample. Those predictions can be provided in two different ways:

- concatenation of these probabilities for each cross-validation iteration,
- determination of probabilities for preselected test dataset values.

The confusion matrix on the other hand shows the normalized number (over the total number of

samples) of predicted values of each class for each class.

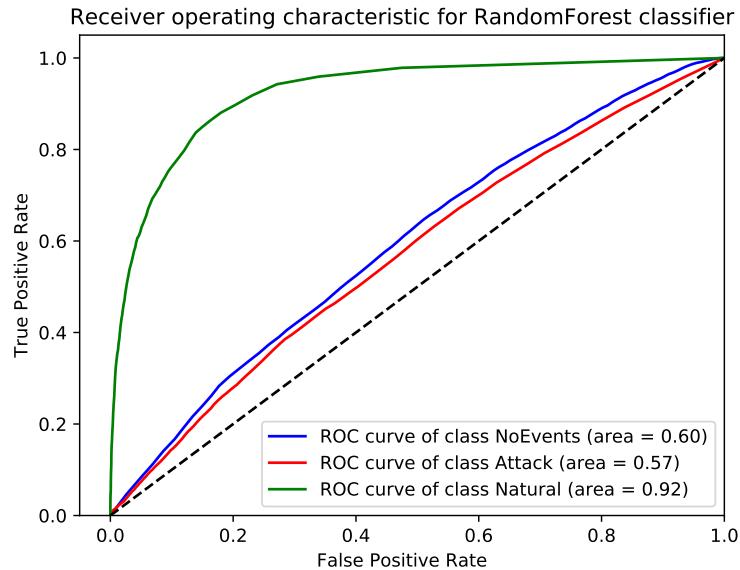
The figure 2-9 represents the ROC curves for the two mentioned types of predictions for Random Forest classifier taught with the power system 3-class dataset. The ROC curves are presented for each class separately and also, as the micro and macro averaged values for the ROC curve calculated based on a preselected test dataset. For the cross validation case ROC curves are almost flattened for NoEvents and Attack classes, whereas for Natural class and all the ROC curves created based on a preselected test data show a hyperbolic shape, which indicates the capacity of the classifier to correctly predict the values in those cases. Figure 2-10 shows the ROC curves plots for SVM classifier. All the ROC curves are practically flat, which means that SVM classifier did not succeed in correctly differentiating between the classes. Figure 2-11 shows the ROC curves for Naïve Bayes classifier, and once more the curves are flat and the classifier did not differentiate between the classes. Finally, figure 2-12 shows the curves for MLP classifier. The curves are generally flat here, however the Attack ROC curve shows a slightly a parabolic shape. In this case, the classifier tends to predict Attack class.

The figure 2-13 shows the confusion matrix for Random Forest classifier taught on 3-class power system dataset using cross-validation. The numbers on the plot indicates that the system tends to predict NoEvents class. The figure 2-14 shows the confusion matrix for SVM classifier. The plot indicates that NoEvents class events tend to be predicted as Attack class and that the classifier does not predict the Natural class at all. The figure 2-15 shows the confusion matrix for Naïve Bayes classifier. In this case the model tends to predict the events as of Natural class. Finally the figure 2-16 for MLP classifier, which tends to predict the NoEvents class.

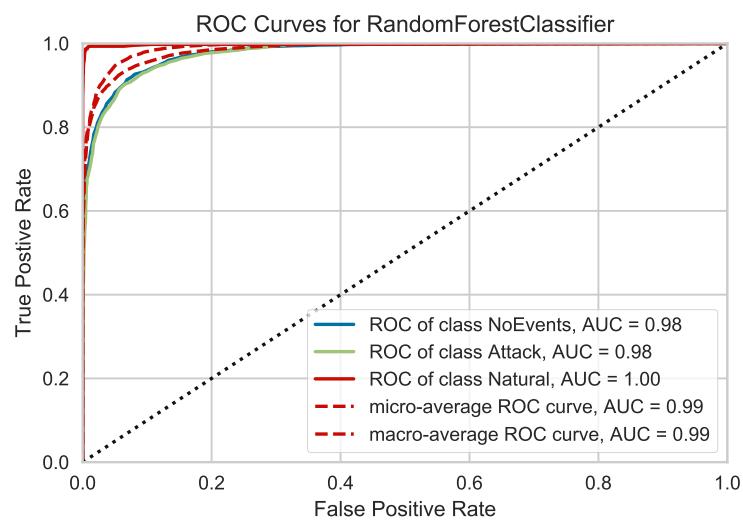
To summarize, the figures show that Random Forest classifier has the higher capacities to distinguish between the occurrence of each class, or its absence. SVM tends to predict only NoEvents and Attacks but does not really succeed in distinguishing between them. Naïve Bayes fails to make true predictions, it considers everything of class natural. Finally MLP, it succeeds in determining the class NoEvents, but does not distinguish over classes almost at all, despite the high accuracy (it is due because of the huge number of samples of class NoEvents).

The two methods of plotting the ROC curve does not give exactly the same results, however the conclusions remain the same in both cases. This fact may be due to a slightly different set of test data used for its creation.

Given this analysis, it can be deducted that Random Forest algorithm acts the best, and that is why it will be adapted in next chapters, in which, at first, an analysis of features and their importance will be made. However a deeper look at the amelioration of will be also made, in order to succeed in differentiation between the classes.

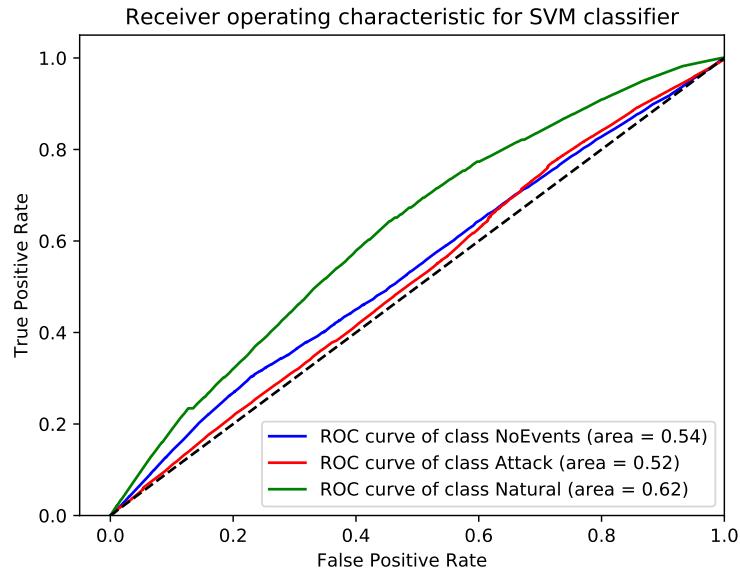


(a) ROC Curve - cross-validation

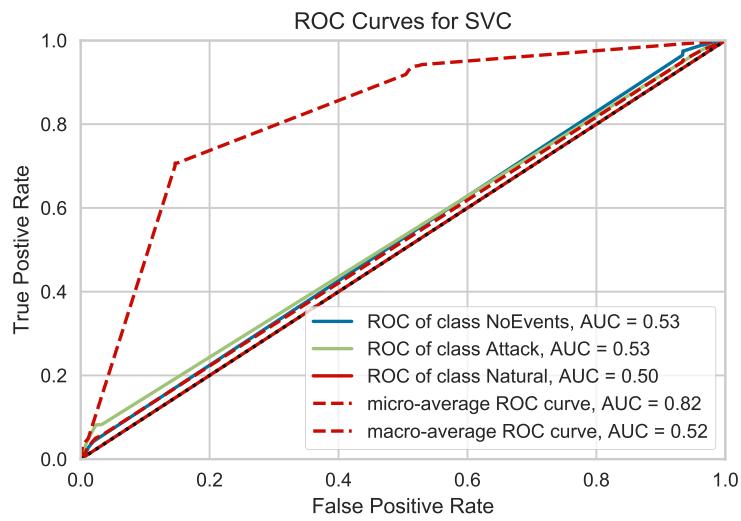


(b) ROC Curve

Figure 2-9: ROC curves plot for Random Forest classifier taught with the power system dataset

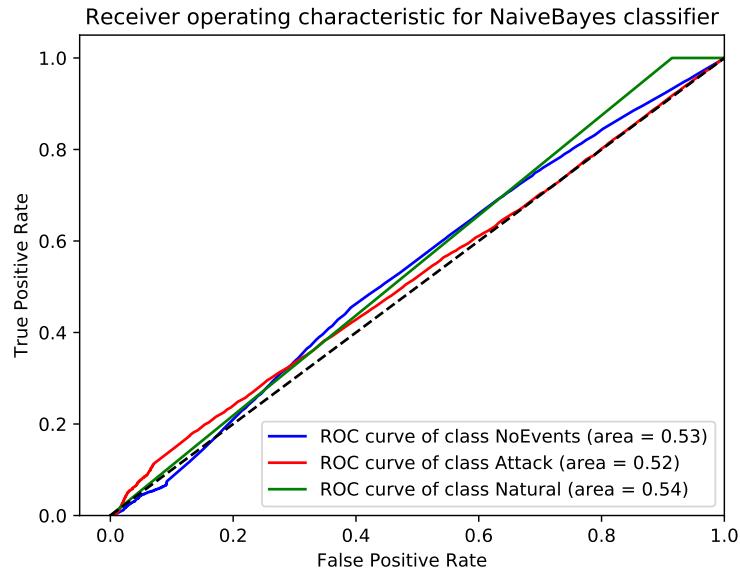


(a) ROC Curve - cross-validation

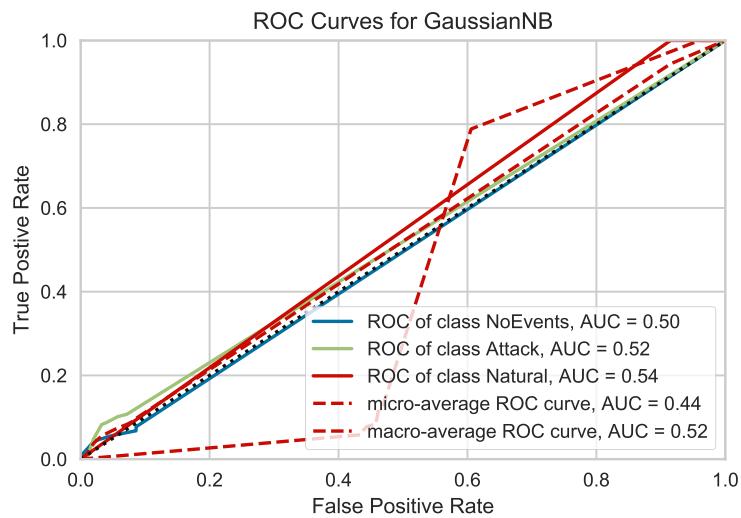


(b) ROC Curve

Figure 2-10: ROC curves plot for SVM classifier taught with the power system dataset

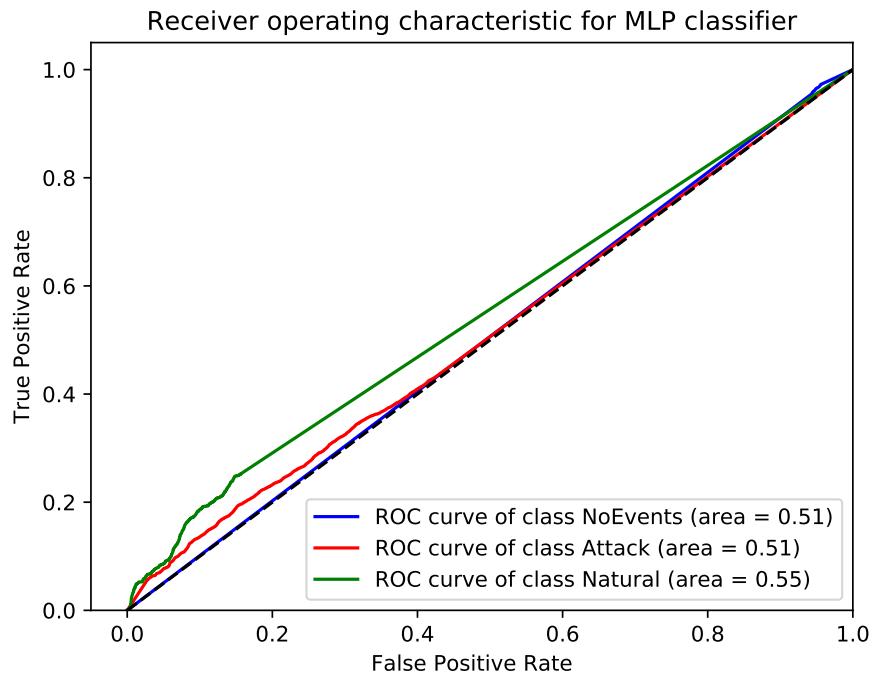


(a) ROC Curve - cross-validation

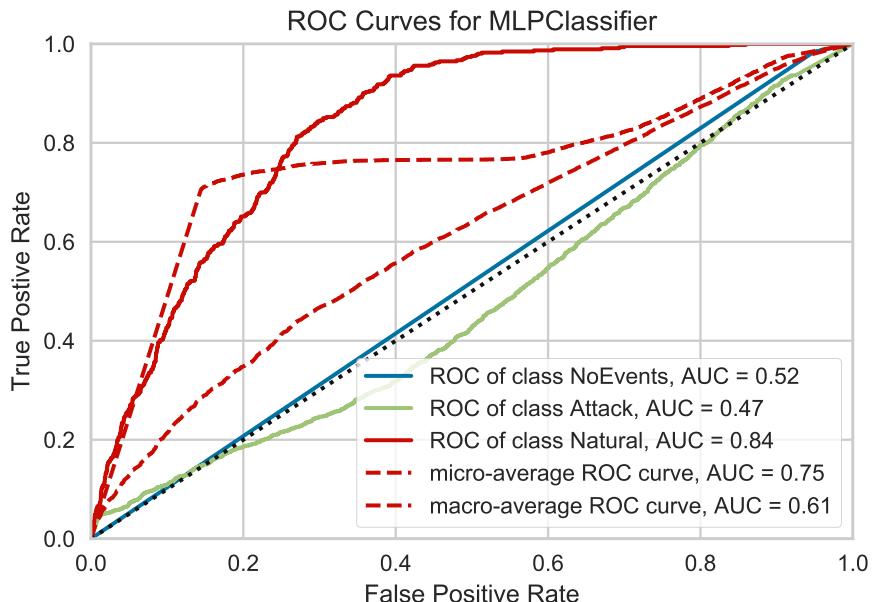


(b) ROC Curve

Figure 2-11: ROC curves plot for Naïve Bayes classifier taught with the power system dataset



(a) ROC Curve - cross-validation



(b) ROC Curve

Figure 2-12: ROC curves plot for MLP classifier taught with the power system dataset

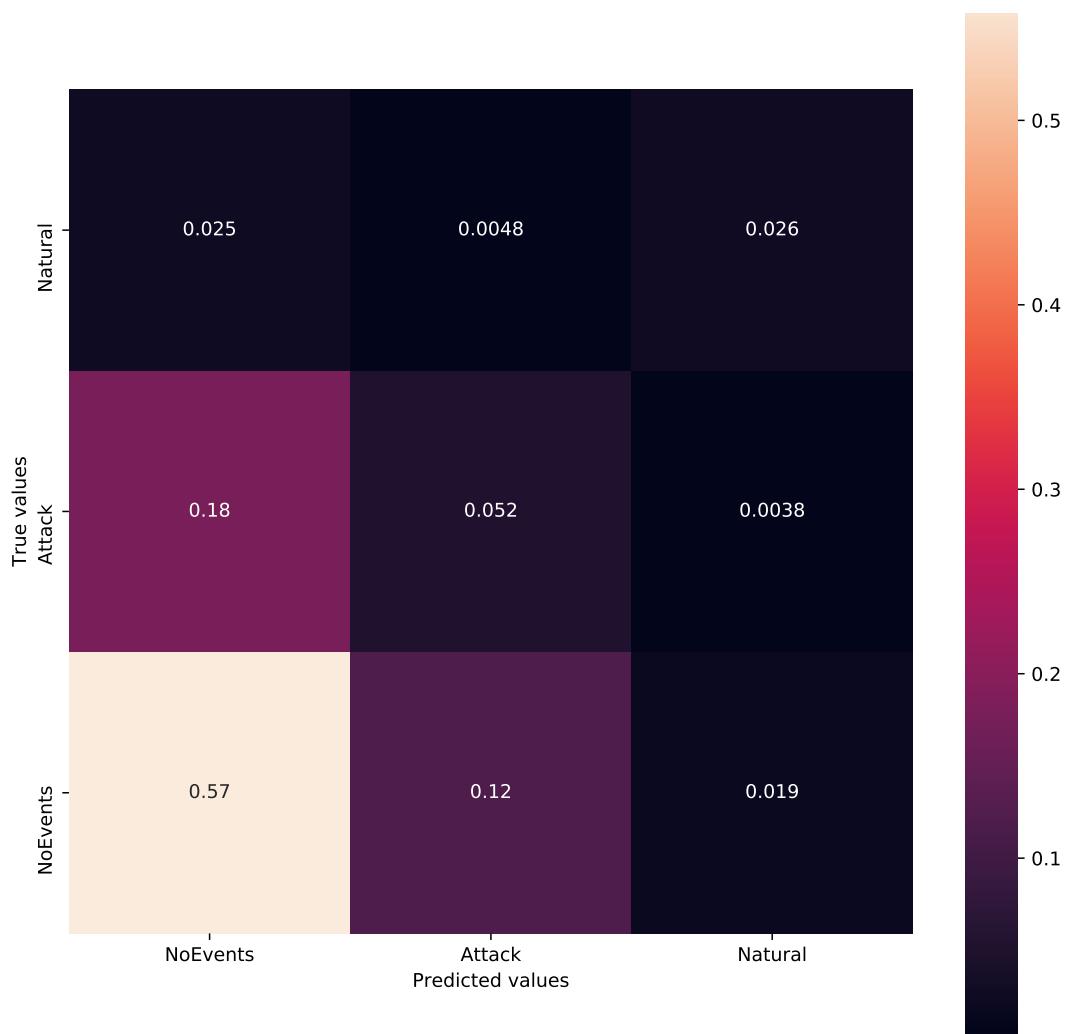


Figure 2-13: Confusion Matrix for Random Forest classifier taught with the power system dataset

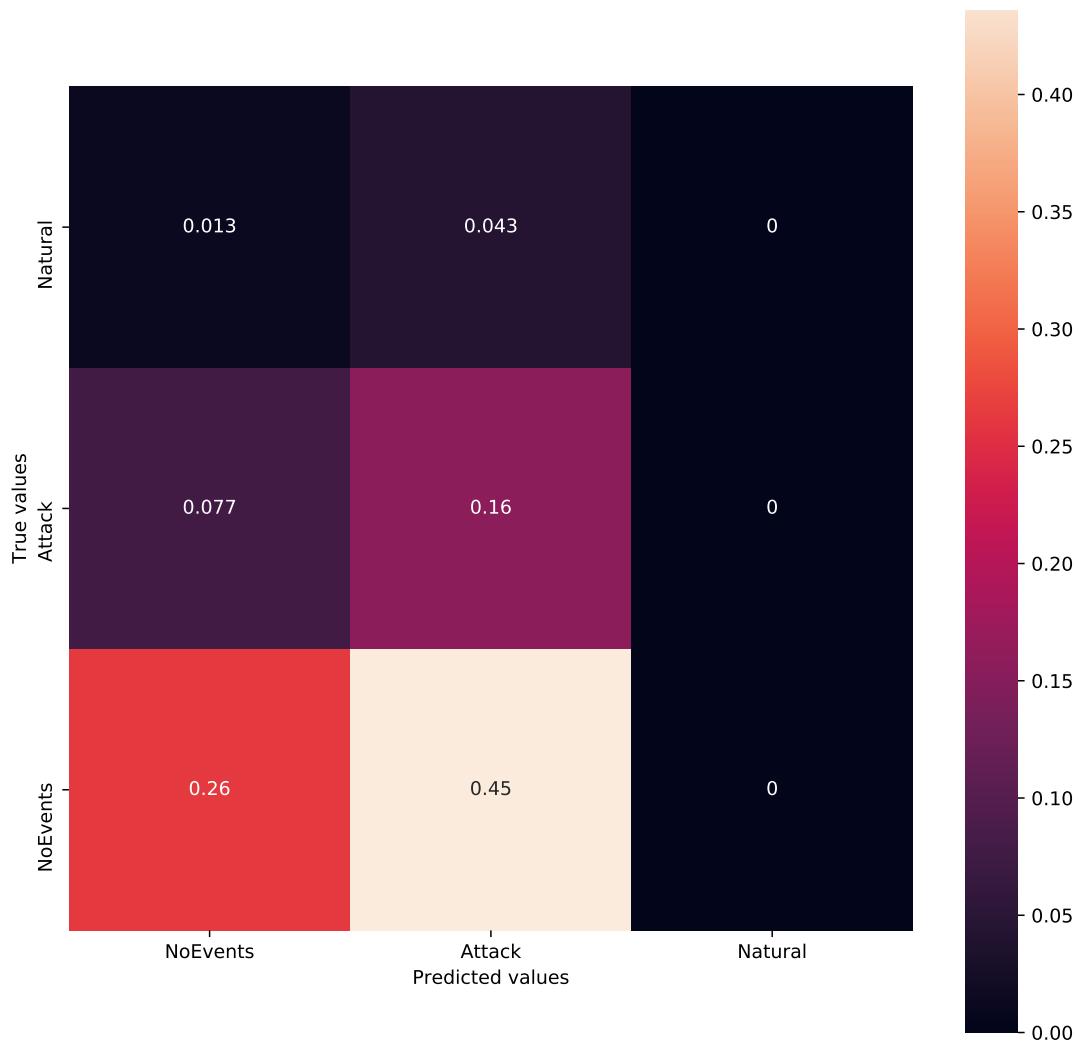


Figure 2-14: Confusion Matrix for SVM classifier taught with the power system dataset

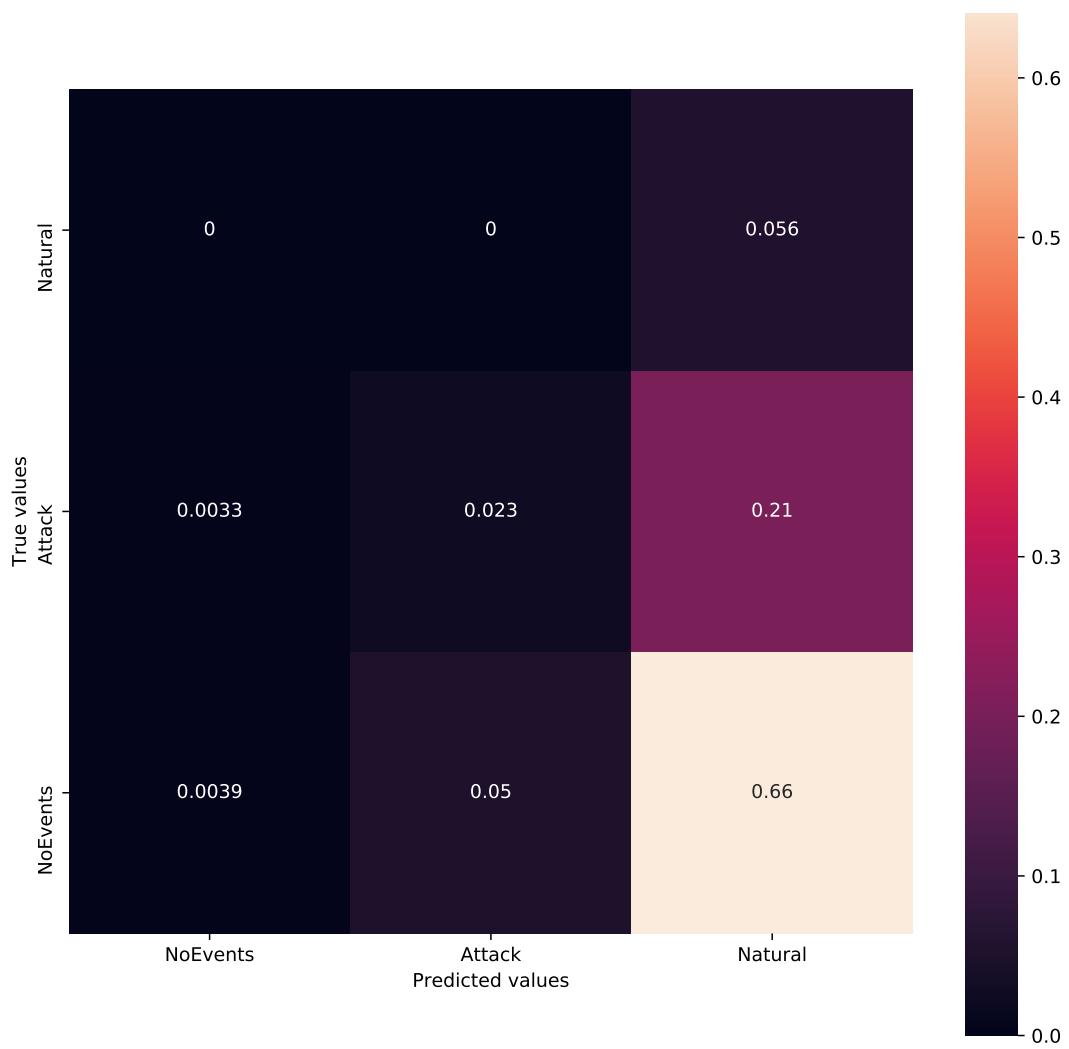


Figure 2-15: Confusion Matrix for Naïve Bayes classifier taught with the power system dataset

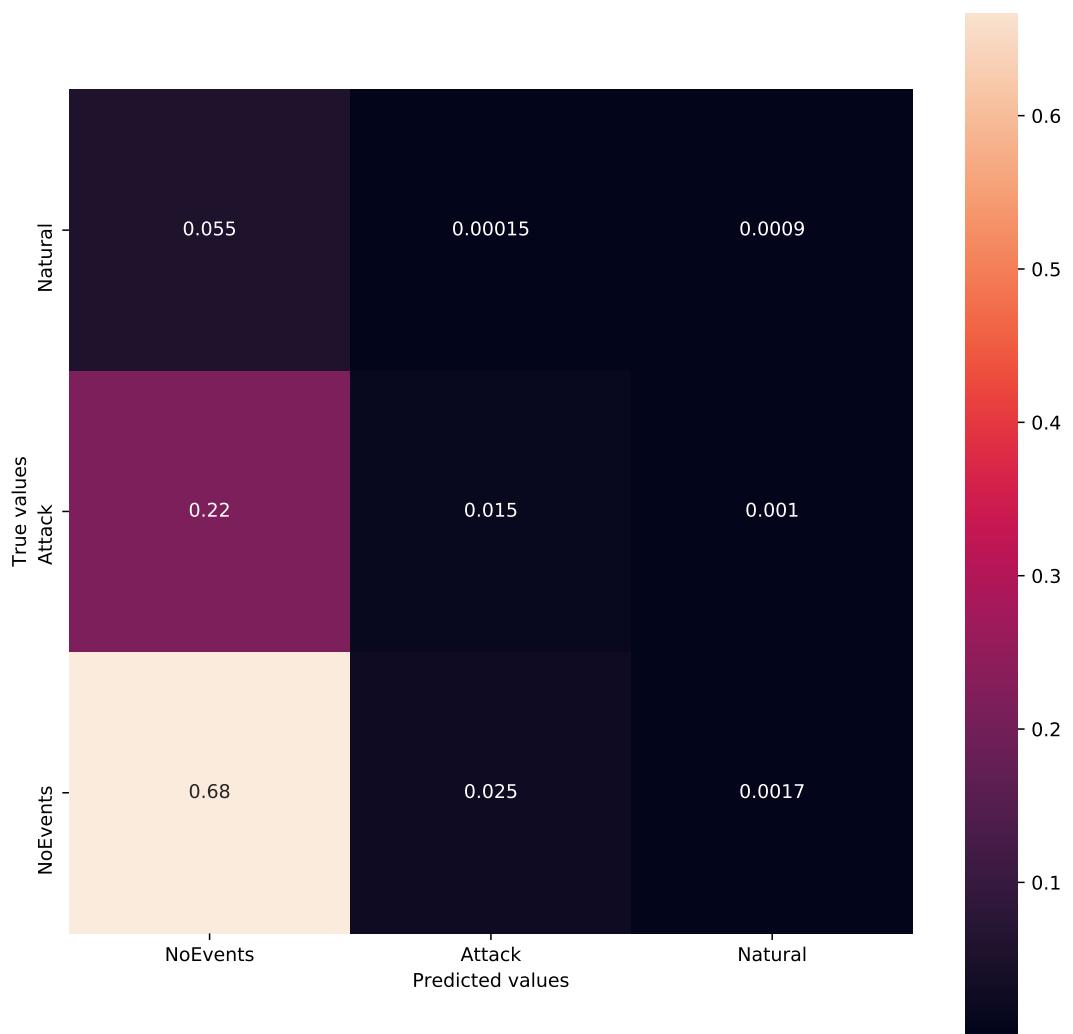


Figure 2-16: Confusion Matrix for MLP classifier taught with the power system dataset

This chapter was a presentation of two machine learning toolkits, four different machine learning techniques and four metrics for the evaluations of those techniques. The values of those metrics were presented for all the techniques using the two toolkits, then compared between themselves and with the results provided by [11]. This permitted to determine the most accurate machine learning technique (Random Forest) which will be used in the next chapters. The next step in this master thesis is the analysis of the importance of features, which will help in further modification of the random forest classifier.

Chapter 3

Features' importance

After having determined the most performant algorithm, which is Random Forest, it is time to go further and analyse which features impact the results of classification the most. The focus will be especially on false predictions. In order to do that, six different tools for feature analysis will be discussed and compared.

3.1 Result's interpreters' comparison

The six mentioned tools are: LIME [30], ELI5 [31], YellowBrick [32], Treeinterpreter [33], dtreeviz [34] and export_graphviz tool from scikit-learn, where the last three ones are designed for Decision Tree and Random Forest classifiers.

3.1.1 LIME

LIME (Local Interpretable Model-agnostic Explanations) is a tool that is used to explain the behaviour of machine learning classifiers. It supports, as for this day, only the explanation of individual predictions for any scikit-learn classifier or regressor. This explanation consist in a list of features (and more precisely decision nodes composed of feature name and a numeric condition) ordered by their relative importance for a particular prediction. This list can be shown in a raw mode (as a python list) or in a visual form (pyplot figure, jupyter notebook or html file).

In order to class the features according to their importance, LIME approximates the model by an interpretable one, created based on perturbing the features of the examined instance. More the perturbed instances are similar to the examined instance, higher is the weight of the perturbed feature. This approximation process is called explanation and is described using the following

equation:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \mathcal{L}(f, g, \pi_x) + \Omega(g) \quad [30], \quad (3.1)$$

where g is the explanation model from the set of interpretable models G , f is the probability that the sample x corresponds to a certain class, π_x is the distance between an instance z and the sample x . $\mathcal{L}(f, g, \pi_x)$ describes to which extent g is unfaithful in the approximation of f in the neighbourhood defined by π_x , while $\Omega(g)$ describes the complexity of the explanation (for example the depth of the tree if f is a decision tree). The goal of the explanation is to find the most faithful approximation of f taking into consideration the neighbourhood π_x .

To better illustrate the algorithm, let the pigeon image in figure 3-1 be considered as a sample used in a classification problem based on a classifier noted f . The image can be divided into several regions, like it is shown in figure 3-2, each region corresponds to a feature.



Figure 3-1: Pigeon



Figure 3-2: Image grid

In the next step, a new dataset is created by modifying the features. In the case of the pigeon image some regions can be simply greyed out like it is shown in figure 3-3. For each of these regions the probability if it represents a pigeon or not is calculated using the classifier f .



(a) Perturbation 1



(b) Perturbation 2



(c) Perturbation 3

Figure 3-3: Greyed-out images forming a new dataset

Using this data it is possible to determine the importance of regions/features of this particular sample. This importance is local and has a meaning only in relation to the sample image from figure 3-1.

3.1.2 ELI5

ELI5 (Explain like I am a 5-year old) is a tool, in form of a Python package, used to debug machine learning classifiers and explain their predictions. It supports multiple machine learning frameworks, including scikit-learn. It can be used to explain how the model works both locally for one prediction and globally for the whole model. The output can take several forms just like in LIME case.

For white-box models, ELI5 works as an extension of scikit-learn and it is capable to extract the weights of model's features for different classes. In addition to that it can show the weights that contributed in a particular prediction. It is done by removing particular features and checking how the accuracy of the classifier was altered. On the other hand, for black-box models, this tool integrates a modified version of LIME, supporting more machine learning frameworks, and a permutation importance method, which checks how the model's accuracy decreases when removing one of the features and on this basis determines the importance of the features.

3.1.3 YellowBrick

YellowBrick is another Python package, which is an extension of scikit-learn framework. It is meant to give global interpretation of the analysed model on different levels. It is possible not only to visualize features importances calculated directly by scikit-learn, but also to give a classification report (accuracy, recall, precision, f-measure), plot a confusion matrix, a ROC curve and much more. In addition to all that, YellowBrick comes with a tool for determination of correlation between features in the dataset.

3.1.4 Treeinterpreter

Treeinterpreter is a simple Python package that works with scikit-learn trees and random forest classifiers. Its only usage consists in decomposing the obtained prediction into bias and contributions of different features. The output is given in the form of a numpy array.

3.1.5 scikit-learn export_graphviz

export_graphviz is a scikit-learn embedded function that enables the user to visualize a decision tree with all the branches and save it into Graphviz¹ format, that can be converted into a vector graphic. It is possible also to visualize decision trees composing random forest model in scikit-learn, since the possibility to extract particular decision trees when using this framework. However the interpretability of results for random forest classifier can be hard.

¹a set of tools for diagram creation using graphs.

3.1.6 dtreeviz

dtreeviz is a more advanced version of export_graphviz available in scikit-learn. For every leaf in the tree it can show a histogram indicating the influence of feature's value on class selection. In addition to that, dtreeviz enables the user to show the path of a particular prediction. The result is saved in the form of a svg vector graphic.

3.1.7 Summary

It can be concluded that ELI5 is the most versatile package compared to others, especially because it enables both global and local interpretations and does not limit its support to scikit-learn, plus it has LIME integrated in it. YellowBrick, on other hand, adds the possibility to analyse from a statistical view the features available in the dataset. Finally comes Treeinterpreter and dtreeviz that are interesting tools when analysing especially Decision Trees. A summary of the most import features of all 5 packages is shown in table 3.1, where 6 comparison metrics where taken into consideration:

- global interpretation: capacity of the tool to interpret the whole model,
- local interpretation: capacity of the tool to interpret a particular sample from the dataset,
- black-box models support: the fact if the tool supports only black-box models (models that can not be simply interpreted),
- features' statistical analysis: the fact if the tool supports statistical analysis of features in the dataset, without taking into consideration the model,
- works only with scikit-learn,
- decision trees graphical visualization.

Table 3.1: Comparison of tools for model analysis

	LIME	ELI5	YellowBrick	Treeinterpreter	dtreeviz
Global interpretation	✗	✓	✓	✗	✗
Local interpretation	✓	✓	✗	✓	✓
Black-box models support	✓	✓	✓	✗	✗
Features' statistical analysis	✗	✗	✓	✗	✗
Works only with scikit-learn	✓	✗	✓	✓	✓
Decision Trees graphical visualization	✗	✗	✗	✗	✓

3.2 Features' importance determination

For the rest of the chapter LIME was chosen to determine the features' importance because of it working with all black-box models available in scikit-learn. The capabilities of LIME were sufficient and that is why ELI5 was not used in his place.

In what follows the Decision Tree classifier will be used because its ease of interpretation. The explication of the algorithm was provided in section 2.2.1 as the part of Random Forest classifier algorithm, since the Random Forest is composed of a certain number of Decision Trees. Moreover the results for Random Forest and Multilayer Perceptron will be presented. The parameters passed to the classifiers are the same as listed in previous chapter.

Because in this case the explanations of single samples are not really interesting, an attempt to generalize the results was made: Lime explainer was run on 100 false predictions of a chosen class. The results are concatenated together. For all the duplicates of features, the mean importance is calculated and only one entry is kept along with this calculated mean value. The python code for the described operations is listed in the listing 3.1.

```
1 import pandas as pd
2 import numpy as np
3 from lime.lime_tabular import LimeTabularExplainer
4
5 # explainer initialization with training data
6 explainer = LimeTabularExplainer(X, training_labels = y, feature_names = features, class_names =
    labels)
7
8 # Xall = all samples, yall = all true classes corresponding to samples (not used in training)
9 boole = (yall != clf.predict(Xall)) # all false predictions
10 faulty = boole & (yall == 0) # false predictions of true class equal 0 (NoEvents)
11 X_test = Xall[faulty] # all false prediction samples
12 y_test = yall[faulty] # all false prediction true classes
13 lst = [] # initialization of a list
14 for idx in range(0, 100): # create explanation for 100 first samples
15     exp = explainer.explain_instance(X_test.iloc[idx], clf.predict_proba, num_features=128,
        labels=[0, 1, 2])
16     lst.append(exp.as_list(label=0))
17 # transform to dataframe
18 lst = np.array(lst)
19 clst = np.concatenate(lst, axis=0)
20 dtfr = pd.DataFrame(clst, columns=['feature', 'importance'])
21 dtfr['importance'] = pd.to_numeric(dtfr['importance'])
22 # group by feature then sort
23 dtfr = dtfr.groupby(['feature']).mean()
24 res = dtfr.sort_values(by='importance')
```

Listing 3.1: Algorithm for determining the most important features using Lime that led to false predictions for 100 samples corresponding to NoEvents class

The results, reduced to 5 most important features for NoEvents, Attack, Natural and all samples, are shown for Decision Tree, Random Forest and MLP classifiers respectively in tables 3.2, 3.3 and 3.4.

The table 3.2 shows the 5 most important features that affected the predictions of the decision tree classifier. The table presents three sub-tables, each of them shows the most important features sorted from the most important to the less important that made the decision tree classifier predict an unexpected class. Each sub-table concerns one of the three classes (NoEvents, Attack and Natural). The table 3.3 represents exactly the same thing, but this time for Random Forest classifier. Finally comes the table 3.4, which concerns the MLP classifier.

Table 3.2: 5 most important features by class for Decision Tree classifier

(a) NoEvents samples		(b) Attack samples	
feature	importance	feature	importance
R4-PA5:IH > 115.38	0.008534	R3:S > 0.00	0.010717
R3-PM2:V <= 128425.29	0.007998	R2-PA7:VH <= -101.20	0.014018
128762.21 < R2-PM1:V <= 129859.49	0.007776	R2-PM1:V > 130872.03	0.009242
0.00 < R1-PA12:IH <= 32.04	0.007514	R3-PA7:VH <= -101.22	0.008676
R3-PM5:I <= 330.70	0.005762	R3-PA2:VH <= -93.75	0.008307
...

(c) Natural samples	
feature	importance
R2:F > 60.00	0.005168
R3:F > 60.00	0.004873
R2-PA5:IH > 63.30	0.004103
R2-PM7:V > 130857.40	0.003693
R1-PA1:VH > 71.28	0.003518
...	...

Table 3.3: 5 most important features by class for Random Forest classifier

(a) NoEvents samples		(b) Attack samples	
feature	importance	feature	importance
R4-PA2:VH > 117.68	0.010149	-97.43 < R1-PA7:VH <= -35.85	0.014091
R3-PM2:V <= 128425.29	0.007994	R4-PA2:VH > 117.68	0.010962
R1-PA:Z > 12.43	0.006171	-97.40 < R1-PA1:VH <= -35.86	0.010908
R2-PM4:I <= 320.95	0.006062	R2-PA6:IH <= -114.38	0.010825
R4-PA4:IH <= -97.95	0.005985	R2-PA2:VH > 114.00	0.009784
...

(c) Natural samples	
feature	importance
R3-PM2:V <= 128425.29	0.007478
R1-PA:Z > 12.43	0.006621
R2-PM4:I <= 320.95	0.006481
R3-PA7:VH > 65.92	0.005920
-62.63 < R3-PA10:IH <= 34.66	0.005893
...	...

Table 3.4: 5 most important features by class for MLP classifier

(a) NoEvents samples		(b) Attack samples	
feature	importance	feature	importance
R4-PM8:V > 0.00	0.111197	R3-PM3:V <= 128676.02	0.084712
R3-PM3:V <= 128676.02	0.082892	R3-PM1:V > 130631.74	0.061883
R3-PM9:V > 0.00	0.066984	129704.03 < R3-PM1:V <= 130631.74	0.052262
R1:S > 0.00	0.053059	128600.80 < R3-PM1:V <= 129704.03	0.041568
R1-PM7:V > 132060.91	0.042774	R3-PM7:V <= 128600.80	0.041327
...

(c) Natural samples	
feature	importance
R3-PM3:V <= 128676.02	0.081114
R3-PM1:V > 130631.74	0.058216
129704.03 < R3-PM1:V <= 130631.74	0.047072
9.59 < R2-PA:Z <= 12.11	0.039845
128600.80 < R3-PM1:V <= 129704.03	0.039490
...	...

It can be observed that the five most important features among false predictions for each class are different for all the classifiers. The only exception is MLP where some features are duplicated, however with different value conditions. This implies creating a correction model for every important feature among all classes. The exact procedure is described in the next chapter.

In this chapter different result interpreters were presented in details. Among them, Lime have been chosen to create an algorithm for features importance determination. The most important features will be used in next chapter to correct the obtained predictions.

Chapter 4

Model enhancement

After having chosen the appropriate machine learning algorithms and found a way to determine the most important features in the classification problem, this chapter is an attempt to create a model able to enhance the results obtained by the basic classification. Three different approaches are presented. First of all, the most important features values were altered and the effect of this modification is examined. Second, a formula for calculating the distance between different samples is established then a way to use that information. Finally, the Hidden Markov Model were used in order to determine the likelihood of prediction of a particular class and this data was used to modify the machine learning algorithm.

4.1 Features values modification

The first proposed solution considers changing the values of most important features in the dataset after the training process. In other words the training process occurs normally, then the importances are determined and a correction function is created (figure 4-1). This correction function will act then on the samples introduced to the model in order to change the features values and obtain better predictions (figure 4-2).

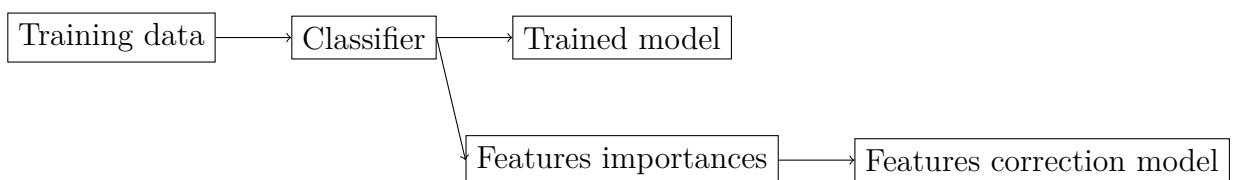


Figure 4-1: Training process illustration



Figure 4-2: Prediction illustration

First of all the correction function was defined as the modification function of the feature values for the predicted samples. This modification consists in shifting the feature value so the feature value would not meet the condition to make a false prediction from tables 3.2-3.4. This function code in Python is shown in listing 4.1.

```
1 def modify(feat, val):
2     X[feat] = X[feat].apply(lambda x: x + val)
```

Listing 4.1: Function used to bulk shift the values of a chosen feature by a given value

where X is a pandas DataFrame object containing all the samples to predict.

Second, the five most important features are taken from tables 3.2-3.4 and the values of features from the samples to predict are altered using the previous function. If a feature is duplicated only the value of the first occurrence is taken into account. The code listed in listing 4.2 shows this operation for Decision Tree classifier.

```
1 modify("R4-PA5:IH", -115.38)
2 modify("R3-PM2:V", 128525.29)
3 modify("R2-PM1:V", 2000)
4 modify("R1-PA12:IH", 32.04)
5 modify("R3-PM5:I", 330.7)
6 modify("R3:S", 0)
7 modify("R2-PA7:VH", 101.20)
8 modify("R2-PM1:V", -1300872.03)
9 modify("R3-PA7:VH", 101.22)
10 modify("R3-PA2:VH", 93.75)
11 modify("R2:F", -60)
12 modify("R3:F", -60)
13 modify("R2-PA5:IH", -63.30)
14 modify("R2-PM7:V", -130857.40)
```

Listing 4.2: Shifting the values of all the most important features using the condition values found using Lime

The samples modified this way are then used for the predictions. In order to check the success of this method, the classification_report method from scikit-learn was used. The results before and after modifying the samples are displayed for Decision Tree, Random Forest and MLP classifiers respectively in tables 4.1, 4.2 and 4.3.

Table 4.1: Features modification results for Decision Tree classifier

(a) Before features modification				(b) After features modification					
	precision	recall	f1-score	support		precision	recall	f1-score	support
NoEvents	0.72	0.76	0.74	51797	NoEvents	0.71	0.83	0.77	51797
Attack	0.27	0.26	0.26	17382	Attack	0.26	0.17	0.21	17382
Natural	0.19	0.08	0.12	4232	Natural	0.10	0.03	0.05	4232
accuracy			0.60	73411	accuracy			0.63	73411
macro avg	0.39	0.37	0.37	73411	macro avg	0.36	0.34	0.34	73411
weighted avg	0.58	0.60	0.59	73411	weighted avg	0.57	0.63	0.59	73411

Table 4.2: Features modification results for Random Forest classifier

(a) Before features modification				(b) After features modification					
	precision	recall	f1-score	support		precision	recall	f1-score	support
NoEvents	0.72	0.87	0.78	51797	NoEvents	0.71	0.89	0.79	51797
Attack	0.29	0.16	0.21	17382	Attack	0.28	0.13	0.18	17382
Natural	0.25	0.05	0.08	4232	Natural	0.26	0.05	0.08	4232
accuracy			0.65	73411	accuracy			0.66	73411
macro avg	0.42	0.36	0.36	73411	macro avg	0.42	0.36	0.35	73411
weighted avg	0.59	0.65	0.61	73411	weighted avg	0.58	0.66	0.61	73411

Table 4.3: Features modification results for MLP classifier

(a) Before features modification				(b) After features modification					
	precision	recall	f1-score	support		precision	recall	f1-score	support
NoEvents	0.71	0.99	0.83	51797	NoEvents	0.74	0.07	0.12	51797
Attack	0.57	0.05	0.09	17382	Attack	0.24	0.93	0.38	17382
Natural	0.00	0.00	0.00	4232	Natural	0.00	0.00	0.00	4232
accuracy			0.71	73411	accuracy			0.27	73411
macro avg	0.43	0.35	0.30	73411	macro avg	0.33	0.33	0.17	73411
weighted avg	0.64	0.71	0.60	73411	weighted avg	0.58	0.27	0.17	73411

The accuracy of the model has slightly increased for both Decision Tree and Random Forest classifiers, and significantly decreased for MLP classifier. Moreover, for Decision Tree classifier, a significant decrease of other metrics for Natural class is observed, for Random Forest some slight increases and decreases of metrics are observed, depending on the class. The most important difference is observed for MLP classifier, where the recall value for NoEvents class passed from 0.99 to 0.07, and in parallel, for Attack class from 0.05 to 0.93.

It may be concluded that this method succeeded to enhance the results for Decision Tree and Random Forest classifiers, but failed with MLP. It is clear that the model, after modifying the features, tends to predict NoEvents class for the first two classifiers, but Attack class for MLP.

4.2 Distance between features

The second proposed solution considers using a transformation routine, which, reduces the number of features to the 15 most important from tables 3.2-3.4 and adds another feature that represents the closest class to the treated sample. The prediction steps were illustrated in figure 4-3.



Figure 4-3: Distance algorithm illustration

The transformation routine on other hand takes the form of a python class and is composed of 4 methods:

1. **distance(X1, X2)**: returns the distance between two samples X1 and X2. It is calculated as the sum of differences between features,
2. **important(X)**: returns the samples with only 15 most important features. The input must be a pandas DataFrame or Series. The choice of features to keep is made by hand directly in the method, without the possibility to change them afterwards,
3. **fit(X,y)**: determines the reference class sample for each class by calculating the mean value of each feature among all samples corresponding to the treated class. It is called only during data fitting to the classifier,
4. **transform(X)**: determines the class with the smallest distance to the sample, based on reference samples determined by fit(X,y) method. The obtained values are then added as a new feature to the samples X and returned afterwards by the method.

This routine has been coupled with the 3 chosen classifiers using **pipeline** class in scikit-learn, which construct acts like a classifier (it has fit and predict methods).

The success rate of this method was verified, once more, using `classification_report` method from scikit-learn. The result and after using the described method are shown for Decision Tree, Random Forest and MLP classifiers respectively in tables 4.4, 4.5 and 4.6.

Table 4.4: Distance routine results for Decision Tree classifier

(a) Before using the distance routine				(b) After using the distance routine					
	precision	recall	f1-score	support		precision	recall	f1-score	support
NoEvents	0.72	0.76	0.74	51797	NoEvents	0.72	0.79	0.75	51797
Attack	0.27	0.26	0.26	17382	Attack	0.28	0.23	0.25	17382
Natural	0.19	0.08	0.12	4232	Natural	0.19	0.08	0.11	4232
accuracy			0.60	73411	accuracy			0.62	73411
macro avg	0.39	0.37	0.37	73411	macro avg	0.40	0.37	0.37	73411
weighted avg	0.58	0.60	0.59	73411	weighted avg	0.58	0.62	0.60	73411

Table 4.5: Distance routine results for Random Forest classifier

(a) Before using the distance routine				(b) After using the distance routine					
	precision	recall	f1-score	support		precision	recall	f1-score	support
NoEvents	0.72	0.87	0.78	51797	NoEvents	0.71	0.83	0.77	51797
Attack	0.29	0.16	0.21	17382	Attack	0.27	0.19	0.22	17382
Natural	0.25	0.05	0.08	4232	Natural	0.18	0.05	0.07	4232
accuracy			0.65	73411	accuracy			0.63	73411
macro avg	0.42	0.36	0.36	73411	macro avg	0.39	0.36	0.36	73411
weighted avg	0.59	0.65	0.61	73411	weighted avg	0.58	0.63	0.60	73411

Table 4.6: Distance routine results for MLP classifier

(a) Before using the distance routine				(b) After using the distance routine					
	precision	recall	f1-score	support		precision	recall	f1-score	support
NoEvents	0.71	0.99	0.83	51797	NoEvents	0.70	0.98	0.82	51797
Attack	0.57	0.05	0.09	17382	Attack	0.18	0.01	0.02	17382
Natural	0.00	0.00	0.00	4232	Natural	0.00	0.00	0.00	4232
accuracy			0.71	73411	accuracy			0.69	73411
macro avg	0.43	0.35	0.30	73411	macro avg	0.29	0.33	0.28	73411
weighted avg	0.64	0.71	0.60	73411	weighted avg	0.54	0.69	0.58	73411

The tables 4.4-4.6 show an increase of accuracy by 0.02 in the Decision Tree classifier case, but a decrease by also 0.02 in the Random Forest and MLP classifiers case. For Decision Tree classifier, the other metrics shows a slight increase, except recall and f-measure for Attack class, however this decrease was not significant. On other hand for the two remaining classifiers a decrease in values of almost all the metrics is observed.

It can be concluded that this method does work only using Decision Tree classifier and gives results better than the features' modification method. For Random Forest and MLP classifiers this method fails to enhance the predictions.

4.3 Hidden Markov Model

The last proposed solution consists in determining the probabilities of occurrence of different classes using so called Hidden Markov Model. The obtained probabilities are then used to modify the split condition in the decision tree classifier.

The Hidden Markov Model is a probabilistic model for generating observable data in a random way by a sequence of internal hidden states, which can not be observed directly. The transitions between those hidden states have the form of a Markov chain [35]. The Markov chain, on other hand, is a chain of interconnected states, where each state's probability depends only of the probability of predeccesing state, without taking into consideration what happened before [36].

In the studied example, the observable data are the features values available in the samples and the states are represented by the 3 classes NoEvents, Attack and Normal. The Hidden Markov Model was used to determine the stationary distribution of the states (invariable in time probability distribution in the Markov chain).

It exists a Python package that implements the Hidden Markov Models and it is compatible with scikit-learn, it's name is **hmmlearn** [37]. It can be used just like any other scikit-learn classifiers, especially that it offers methods like `fit()`, `predict()`, `predict_proba()`, etc...

Using **hmmlearn** the stationary distribution was determined in 4 lines of code as shown in listing 4.3.

```

1 from hmmlearn.hmm import GaussianHMM
2 clf2 = GaussianHMM(n_components = 3) # 3 states
3 clf2 . fit (X)
4 coefs = clf2 . get_stationary_distribution ()
```

Listing 4.3: Determination of the stationary distribution of the classes using hmmlearn

The Gaussian HMM was used because the observations (features) are continuous. It was initialized by a number of components equal to 3 because of 3 states - NoEvents, Attack and Natural.

The stationary distribution obtained this way was then used to set the `class_weight` parameter in `DecisionTreeClassifier` in scikit-learn, as shown in the listing 4.4.

```
1 ctest = DecisionTreeClassifier(class_weight = {0: coefs [0] , 1:coefs [1] , 2:coefs [2]})
```

Listing 4.4: Definition of the Decision Tree Classifier taking into consideration the stationary distribution

For Random Forest classifier the same parameter exists and was used to produce results. However, for MLP classifier a similar parameter does not exist and using the determined coefficients would require deeper modifications of the neural network.

The results for Decision Tree and Random Forest classifiers are presented in tables 4.7 and 4.8.

Table 4.7: Hidden Markov Model manipulation results for Decision Tree classifier

(a) Before using HMM				(b) After using HMM					
	precision	recall	f1-score		precision	recall	f1-score	support	
NoEvents	0.72	0.76	0.74	51797	NoEvents	0.71	0.75	0.73	51797
Attack	0.27	0.26	0.26	17382	Attack	0.27	0.26	0.26	17382
Natural	0.19	0.08	0.12	4232	Natural	0.16	0.06	0.09	4232
accuracy			0.60	73411	accuracy			0.60	73411
macro avg	0.39	0.37	0.37	73411	macro avg	0.38	0.36	0.36	73411
weighted avg	0.58	0.60	0.59	73411	weighted avg	0.58	0.60	0.59	73411

Table 4.8: Hidden Markov Model manipulation results for Random Forest classifier

(a) Before using HMM				(b) After using HMM					
	precision	recall	f1-score		precision	recall	f1-score	support	
NoEvents	0.72	0.87	0.78	51797	NoEvents	0.72	0.85	0.78	51797
Attack	0.29	0.16	0.21	17382	Attack	0.30	0.19	0.24	17382
Natural	0.25	0.05	0.08	4232	Natural	0.25	0.05	0.09	4232
accuracy			0.65	73411	accuracy			0.65	73411
macro avg	0.42	0.36	0.36	73411	macro avg	0.42	0.37	0.37	73411
weighted avg	0.59	0.65	0.61	73411	weighted avg	0.59	0.65	0.61	73411

The tables 4.7 and 4.8 show that the accuracy did not change after using the probabilities obtained from Hidden Markov Model. Looking more in details, it can be observed a general small decrease of the majority of the other metrics in the Decision Tree classifier case, but, in Random Forest classifier case, a slight increase in metrics value in relation to the Attack class and a small decrease of recall for NoEvents.

It may be concluded that this method does not work with the given dataset, however the results obtained by the Random Forest classifier are better than those provided by Decision Tree classifier. The possible cause of the failure of this method may be the fact that the samples in the dataset does not include any information about their order in time. If the dataset contained information about the time of the event, the results, probably, would be more significant.

In this chapter, three methods for results' enhancement were presented. The results vary depending on the used machine learning technique. For Decision Tree classifier, the best method for enhancing the results was the second, using the distance between features. For Random Forest classifier, the best method was the first, modifying the values of features. Finally, for MLP classifier, both two tested methods failed, but the second did not decrease the metrics that drastically as the first method.

Conclusions

The aim of this thesis was to develop a hybrid neural network capable of differentiating between three states of a cyber physical system - normal behaviour, a fault or an attack. To facilitate the work and focus on computer science topics, ready datasets for a power system was taken and analysed in chapter 2.

Three different methods were presented to achieve the initial goal - modification of values of features, method using distances between samples and a method using Hidden Markov Model. They were tested on three different machine learning classifiers - Decision Tree, Random Forest and Multilayer Perceptron.

For Decision Tree classifier the most efficient enhancement method was using the proposed method that works on distances between samples. For Random Forest classifier, the most effective method was modifying the values of features. And for Multilayer Perceptron classifier all the proposed methods failed.

Hidden Markov Model method gave the smallest impact on the tested classifiers. This can be explained by the time independence of the used dataset, especially that this method relies on order of samples.

A possible better solution for the posed problem may be using a physical model of the power system that could estimate on its own some features and combining it with a machine learning technique. There exist some python toolkits created to model a power system like PyPSA [38] or pandapower [39], but they are complex and require specific advanced technical skills.

This thesis in addition to all that, provides some other important conclusions. First, the different machine learning toolkits do not work always the same way and the results may differ from one toolkit to another, like in this case and the differences in results between Weka and scikit-learn. Second, in order to the results from a work to be reproducible, all the details concerning the used parameters and the version of the software must be provided, in other case it may be hard to get the same result. Finally, it exists plenty of tools for machine learning classifiers evaluation and each toolkit provides its own interpretation of results.

This work may be continued by implementing the power system using one of the mentioned

toolkits in order to try to further enhance the results. In addition to that, instead of using scikit-learn, Keras [27] or PyTorch [26] packages may be used in order to create a complex neural network and possibly get better results.

List of Figures

0-1	Physical features addition algorithm visualization	7
0-2	Attack generation approach training algorithm visualization	7
1-1	Power system network diagram	10
1-2	Scenarios distribution throughout all 15 files	12
1-3	Scenarios distribution throughout the 3-class dataset files	12
1-4	Magnitude and angle (θ) of sinusoidal functions	13
2-1	Plots showing the accuracy values per dataset file for different machine learning techniques using three-class datasets of the power system calculated using Weka and scikit-learn and compared to results from [11]	26
2-2	Plots showing the accuracy values per dataset file for different machine learning techniques using binary datasets of the power system calculated using Weka and scikit-learn and compared to results from [11]	27
2-3	Plots showing the accuracy values per dataset file for different machine learning techniques using multiclass datasets of the power system calculated using Weka and scikit-learn and compared to results from [11]	28
2-4	Plots showing the precision metric values for different machine learning techniques using different datasets of the power system calculated using Weka and scikit-learn and compared to results from [11]	29
2-5	Plots showing the recall metric values for different machine learning techniques using different datasets of the power system calculated using Weka and scikit-learn and compared to results from [11]	30
2-6	Plots showing the f-measure metric values for different machine learning techniques using different datasets of the power system calculated using Weka and scikit-learn and compared to results from [11]	31
2-7	Plots illustrating F-measure macro average values for different classifiers using different datasets of the power system made using Weka and scikit-learn	33

2-8	Plots illustrating F-measure micro average values for different classifiers using different datasets of the power system made using Weka and scikit-learn	34
2-9	ROC curves plot for Random Forest classifier taught with the power system dataset .	36
2-10	ROC curves plot for SVM classifier taught with the power system dataset	37
2-11	ROC curves plot for Naïve Bayes classifier taught with the power system dataset . . .	38
2-12	ROC curves plot for MLP classifier taught with the power system dataset	39
2-13	Confusion Matrix for Random Forest classifier taught with the power system dataset	40
2-14	Confusion Matrix for SVM classifier taught with the power system dataset	41
2-15	Confusion Matrix for Naïve Bayes classifier taught with the power system dataset . .	42
2-16	Confusion Matrix for MLP classifier taught with the power system dataset	43
3-1	Pigeon	46
3-2	Image grid	46
3-3	Greyed-out images forming a new dataset	46
4-1	Training process illustration	53
4-2	Prediction illustration	53
4-3	Distance algorithm illustration	56

List of Tables

1.1	IED measurements	14
2.1	Random Forest classifier used parameters in Weka and scikit-learn	19
2.2	SVM classifier used parameters in Weka and scikit-learn	20
2.3	NaïveBayes classifier used parameters in Weka and scikit-learn	21
2.4	MLP classifier used parameters in scikit-learn	23
3.1	Comparison of tools for model analysis	48
3.2	5 most important features by class for Decision Tree classifier	50
3.3	5 most important features by class for Random Forest classifier	51
3.4	5 most important features by class for MLP classifier	51
4.1	Features modification results for Decision Tree classifier	55
4.2	Features modification results for Random Forest classifier	55
4.3	Features modification results for MLP classifier	55
4.4	Distance routine results for Decision Tree classifier	57
4.5	Distance routine results for Random Forest classifier	57
4.6	Distance routine results for MLP classifier	57
4.7	Hidden Markov Model manipulation results for Decision Tree classifier	59
4.8	Hidden Markov Model manipulation results for Random Forest classifier	59

List of Listings

3.1	Algorithm for determining the most important features using Lime that led to false predictions for 100 samples corresponding to NoEvents class	49
4.1	Function used to bulk shift the values of a chosen feature by a given value	54
4.2	Shifting the values of all the most important features using the condition values found using Lime	54
4.3	Determination of the stationary distribution of the classes using hmmlearn	58
4.4	Definition of the Decision Tree Classifier taking into consideration the stationary distribution	58

Bibliography

- [1] A. Karpatne, W. Watkins, J. Read, and V. Kumar, “Physics-guided Neural Networks (PGNN): An Application in Lake Temperature Modeling,” *arXiv:1710.11431 [physics, stat]*, Feb. 2018. <http://arxiv.org/abs/1710.11431>.
- [2] E. M. Ferragut, J. Laska, M. M. Olama, and O. Ozmen, “Real-Time Cyber-Physical False Data Attack Detection in Smart Grids Using Neural Networks,” in *2017 International Conference on Computational Science and Computational Intelligence (CSCI)*, (Las Vegas, NV, USA), pp. 1–6, IEEE, Dec. 2017. <https://ieeexplore.ieee.org/document/8560712/>.
- [3] R. Swischuk, L. Mainini, B. Peherstorfer, and K. Willcox, “Projection-based model reduction: Formulations for physics-based machine learning,” *Computers & Fluids*, vol. 179, pp. 704–717, Jan. 2019. <http://www.sciencedirect.com/science/article/pii/S0045793018304250>.
- [4] A. S. Zamzam and N. D. Sidiropoulos, “Physics-Aware Neural Networks for Distribution System State Estimation,” *arXiv:1903.09669 [cs, math]*, July 2019. <http://arxiv.org/abs/1903.09669>.
- [5] A. Karpatne, G. Atluri, J. H. Faghmous, M. Steinbach, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar, “Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, pp. 2318–2331, Oct. 2017. <https://ieeexplore.ieee.org/document/7959606/>.
- [6] M. Hipsey, L. Bruce, and D. Hamilton, “GLM - General Lake Model: Model overview and user information,” *The University of Western Australia, Perth, Australia*, 2014.
- [7] T. Morris, “Industrial Control System (ICS) Cyber Attack Datasets - Tommy Morris.” <https://sites.google.com/a/uah.edu/tommy-morris-uah/ics-data-sets>.
- [8] U. Adhikari, S. Pan, and T. Morris, “Power System Attack Datasets,” Apr. 2014.
- [9] “Three-phase electric power,” *Wikipedia*, May 2020. https://en.wikipedia.org/w/index.php?title=Three-phase_electric_power&oldid=955675515.

- [10] J.L. Kirtley Jr., “Introduction To Symmetrical Components,” *6.061 Introduction to Power Systems*, vol. Class Notes Chapter 4. https://ocw.mit.edu/courses/electrical-engineering-and-computer-science/6-061-introduction-to-electric-power-systems-spring-2011/readings/MIT6_061S11_ch4.pdf.
- [11] R. C. Borges Hink, J. M. Beaver, M. A. Buckner, T. Morris, U. Adhikari, and S. Pan, “Machine learning for power system disturbance and cyber-attack discrimination,” in *2014 7th International Symposium on Resilient Control Systems (ISRCS)*, (Denver, CO, USA), pp. 1–8, IEEE, Aug. 2014. <http://ieeexplore.ieee.org/document/6900095/>.
- [12] “Appendix B - The WEKA workbench,” in *Data Mining (Fourth Edition)* (I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, eds.), pp. 553–571, Morgan Kaufmann, fourth edition ed., 2017. <http://www.sciencedirect.com/science/article/pii/B9780128042915000246>.
- [13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [14] “Who is using scikit-learn? — scikit-learn 0.23.1 documentation.” <https://scikit-learn.org/stable/testimonials/testimonials.html>.
- [15] “RandomForest.” <https://weka.sourceforge.io/doc.dev/weka/classifiers/trees/RandomForest.html>.
- [16] “Sklearn.ensemble.RandomForestClassifier — scikit-learn 0.23.1 documentation.” <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html>.
- [17] A. Yadav, “SUPPORT VECTOR MACHINES(SVM).” <https://towardsdatascience.com/support-vector-machines-svm-c9ef22815589>, Oct. 2018.
- [18] “Support Vector Machines, Clearly Explained!!!.” <https://www.youtube.com/watch?v=efR1C6CvhmE>, Jan. 2014.
- [19] “LibSVM.” <https://javadoc.scijava.org/Weka/weka/classifiers/functions/LibSVM.html>.
- [20] “Sklearn.svm.SVC — scikit-learn 0.23.1 documentation.” <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>.

- [21] “NaiveBayes.” <https://weka.sourceforge.io/doc.dev/weka/classifiers/bayes/NaiveBayes.html>.
- [22] “Sklearn.naive_bayes.GaussianNB — scikit-learn 0.23.1 documentation.” https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.GaussianNB.html.
- [23] V. Patel, “Implementing a Multi Layer Perceptron Neural Network in Python.” <https://towardsdatascience.com/implementing-a-multi-layer-perceptron-neural-network-in-python-b22b5a3bdaf3>, May 2020.
- [24] “Neural network models (supervised).” https://scikit-learn.org/stable/modules/neural_networks_supervised.html.
- [25] K. Stokfiszewski, “Soft Computing Laboratory 1. The Delta rule.”
- [26] “PyTorch.” <https://www.pytorch.org>.
- [27] “Keras: The Python deep learning API.” <https://keras.io/>.
- [28] “Sklearn.neural_network.MLPClassifier — scikit-learn 0.23.1 documentation.” https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html.
- [29] B. Shmueli, “Multi-Class Metrics Made Simple, Part II: The F1-score.” <https://towardsdatascience.com/multi-class-metrics-made-simple-part-ii-the-f1-score-ebe8b2c2ca1>, June 2020.
- [30] M. T. Ribeiro, S. Singh, and C. Guestrin, “"Why Should I Trust You?": Explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 1135–1144, 2016.
- [31] Mikhail Korobov and Konstantin Lopuhin, “ELI5.” <https://github.com/TeamHG-Memex/eli5>.
- [32] B. Bengfort, R. Bilbro, N. Danielsen, L. Gray, K. McIntyre, P. Roman, Z. Poh, *et al.*, “Yellowbrick,” 2018-11-14, 2018. <http://www.scikit-yb.org/en/latest/>.
- [33] Ando Saabas, “TreeInterpreter.” <https://github.com/andosa/treeinterpreter>.
- [34] Terence Parr and Prince Grover, “Dtreetviz.” <https://github.com/parrt/dtreetviz>.

- [35] “Tutorial — hmmlearn 0.2.3.post13+ge872126 documentation.” <https://hmmlearn.readthedocs.io/en/latest/tutorial.html#building-hmm-and-generating-samples>.
- [36] T. Amit, “Introduction to Hidden Markov Models.” <https://towardsdatascience.com/introduction-to-hidden-markov-models-cd2c93e6b781>, Oct. 2019.
- [37] “Hmmlearn.” <https://github.com/hmmlearn/hmmlearn>, Aug. 2020.
- [38] T. Brown, J. Hörsch, Fabian Hofmann, F. Neumann, R. Smith, Chloe, Schlachtberger, Jdedecca, Stianchris, F. V. D. Plas, H. Liu, M. Maria, G. Yu, M. Bilousova, M. Lavoie, M. H. Duong, Nmartensen, P. Belotti, Dcradu, and Jankaeh, “PyPSA/PyPSA: PyPSA Version 0.17.1.” Zenodo, July 2020. <https://zenodo.org/record/3946412>.
- [39] “Pandapower.” <http://www.pandapower.org/>.