UNIVERSITY KASDI MERBAH OUARGLA
FACULTY OF MATHEMATICS AND MATERIAL SCIENCES
DEPARTEMENT OF PHYSICS

MATERIALS PHYSICS

BY :
KHERFI MERIEM

Prediction of Gaunt factor by using Support vector regression

Theses supported on 21 /05 /2025

Before the jury composed of :

Omar Bentouila	University Kasdi Merbah Ouargla	President
Abdelhakim Benkrane	University Kasdi Merbah Ouargla	Examiner
Djamel Eddine Zenkhri	University Kasdi Merbah Ouargla	Supervisor

Abstract

This research aims to study the possibility of using the Support Vector Regression(SVR) algorithm to predict the values of the Gaunt factor, which is a fundamental element in quantum calculations, especially in describing interactions between electrons in multi-electron atoms. This factor is used in complex angular integrals and plays a pivotal role in electron spectral and atomic structure calculations. A predictive model was developed using SVR based on a pre-calculated dataset containing inputs such as angular quantum numbers and interacting orbitals. Several types of kernel functions were tested, and the parameters were tuned to achieve the best predictive performance. The model results showed good accuracy (0.999) and the mean square error RMSE (0.003) in predicting the values of the Gaunt factor, and also demonstrated the algorithm's efficiency in representing complex nonlinear relationships between variables. This research contributes to opening new avenues for accelerating computations in quantum mechanics and confirms the feasibility of integrating machine learning techniques into theoretical sciences, paving the way towards developing more efficient tools for studying complex quantum systems.

Keywords:

Prediction of Gaunt factor, Support vector regression(SVR), Machine learning, Application of SVR in the prediction, Improve the performance of SVR.

ملخص

يهدف هذا البحث إلى دراسة إمكانية استخدام خوارزمية الانحدار المتجه الداعم في التنبؤ بقيم معامل Gaunt, الذي يعد من العناصر الأساسية في الحسابات الكمومية, خصوصا في وصف التآثرات بين الإلكترونات. يستخدم هذا المعامل في التكاملات الزاوية المعقدة و يؤدي دورا محوريا في حسابات الطيف الإلكتروني و البنية الذرية. تم تطوير نموذج تنبؤي باستخدام SVR بالاعتماد على مجموعة بيانات محسوبة مسبقا, تحتوي على مدخلات. تم اختبار عدة أنواع من دوال النواة, وتم ضبط المعلمات باستخدام تقنيات التحقق المتقاطع لتحقيق أفضل أداء تنبؤي. أظهرت النتائج أن النموذج يحقق دقة تنبؤية عالية بلغت 0.999, كما بلغ متوسط مربع الخطأ حوالي 0.003 , وهو ما يدل على كفاءة النموذج في تمثيل العلاقة المعقدة بين المتغيرات الفيزيائية المستقلة وقيم معامل .Gaunt يساهم هذا البحث في فتح آفاق جديدة لتسريع العمليات الحسابية في ميكانيك الكم, و يؤكد على جدوى دمج تقنيات تعلم الآلة في مجالات العلوم النظرية, مما يمهد الطريق نحو تطوير أدوات أكثر كفاءة لدراسة الأنظمة الكمومية المعقدة.

:

الكلمات المفتاحية:

التنبؤ بمعامل جونت, الانحدار المتجه الداعم, التعلم الألي, تطبيق الإنحدار المتجه الداعم في التنبؤ, تحسين أداء الإنحدار المتجه الداعم .

Dedication

To those who were the fortress, the motivation, and the support in every step, to those who gave me the pillars of continuity with their love, and the provisions for the path with their prayers, I dedicate this humble work to those who deserve all appreciation and gratitude.

To my dear father, you who taught me that the path begins with will and is built with patience. Thank you for your steadfastness, from which I drew my strength, and for your words that were a beacon in times of confusion.

To my dear mother, you who were the raised prayer, the caring hand, and the light in the darkness of days, all my love to you, for you are the homeland, the safety, and the secret to every success that has been achieved and will be written.

To my second mother, who embraced me with the love of her heart and accompanied me with her tenderness, thank you for your silent giving and warm presence. You have always been a support, no less than a mother, but rather an extension of her.

To my dear brothers, you are my family that I am proud of; you were the silent supporters and the ones who rejoiced before me at every achievement. Thank you for your pure hearts and steadfast positions.

To all my family, near or far, to everyone who had a sincere invitation, a word of support, or a moment of interest, you are part of this achievement, and even if you did not see it, its effects are between the lines.

This work is a token of gratitude to everyone who left a mark on my heart, and a mark of loyalty to everyone who was part of this journey.

Acknowledgments

All praise and thanks be to Allah, the Almighty, who guided me in accomplishing this work. I extend my sincere gratitude and appreciation to my supervisor, Mr. Djamel Eddine Zenkhri, for his continuous support and guidance throughout my research journey. I also express my heartfelt thanks to the assistant professor, Mr. Benkrane Abdelhakim, for his insightful feedback and encouragement. My appreciation also goes to the president of the jury, Mr. Bentouila Omar, for overseeing the defense process with great professionalism.

I am also thankful to Ms. Ayat Zahia, who provided me with valuable advice and support throughout the preparation of my thesis. I express my utmost gratitude to the members of the jury who dedicated their time and effort to evaluate my work. Your feedback has added significant value and helped improve the quality of my thesis.

I would like to thank all the professors at the Faculty of Material Sciences, Kasdi Merbah University, for their guidance and support during my academic journey. Lastly, I extend my deepest gratitude to everyone who contributed to this thesis, directly or indirectly, and offered assistance in various ways. I am truly grateful for their support, helpful advice, and the optimism they shared with me. To all of them, I extend my sincere thanks and appreciation.

Contents

List of Figures

List of Tables

General Introduction

The Gaunt factor is a pivotal element in quantum mechanics, given its fundamental role in calculations related to angular interactions between electrons within multi-electron atoms refer to a type of electrostatic interaction that arises due the spatial distribution of electrons around the nucleus and their mutual influence based on angle, not just distance. These factors appear in triplet angle integrals associated with spherical harmonic functions, which are widely used to describe electronic interactions with atoms and molecules. These factors contribute directly to the characterization of electronic structure, atomic spectra, and model material properties at the quantum level[7].

Despite the importance of these parameters, calculating the, accurately using methods is often complex and computationally time-consuming, especially when it comes to atomic systems with a large number of electrons. Hence, the need to develop alternative methods that guarantee both efficiency and accuracy emerged, paving the way for the use of artificial intelligence and machine learning techniques as modern tools in supporting quantum research.

In this context, the support vector regression (SVR) algorithm stands out as one of the most prominent machine learning techniques used in forecasting and data analysis, as it has proven effective in handling nonlinear problems even with limited data availability. This research aims to exploit the SVR algorithm to build a predictive model capable of estimating Gaunt factor values with high accuracy, by training the model on data previously calculated using traditional methods.

The Gaunt factor is a mathematical integral involving the product of three harmonic spherical functions. It is widely used in angular momentum coupling calculations in quantum systems. This factor appears clearly in multi-electron atomic calculations, where it is used to evaluate angular integrals that describe interactions between different electrons. This integral expresses the complex relationships between angular quantum numbers and is a fundamental element in atomic theories.

The difficulty of calculating the Gaunt factors increases with the complexity of the system under study, which makes finding efficient and fast alternatives for

accurately estimating these factors a necessity in several fields, such as solid-state physics, theoretical chemistry, and advanced materials design.

Support Vector Regression (SVR) is a machine learning technique based on the principles of the vector machine (SVM) algorithm, and was developed to handle regression problems rather than classification. The idea of SVR is to find a predictive function that achieves the best possible approximation of the relationship between the input, such that the prediction error is within acceptable limits(known as the epsilon margin).

SVR is characterized by its high flexibility compared to other models, such as GBR and ANN, in terms of accuracy in representing nonlinear relationships thanks to the use of (kernel function), which transforms data from the input space into a higher-dimensional space, allowing it to be separated or represented more accurately. The radial kernel (RBF) is one of the most widely used kernels due to its ability to handle diverse data types.

Given these capabilities, the SVR model is expected to be an effective tool in predicting the Gaunt factor, especially when the relationships between variables are complex and nonlinear, as is the case in quantum contexts.

Chapter 1

Gaunt factor

1.1 Introduction:

The Gaunt factor is one of the fundamental elements in quantum calculations related to the electronic structure of atoms, as it plays a pivotal role in determining the values of orbital angular momentum matrices. This factor arises from angular integrals between spherical harmonic functions and is widely used in atomic physics, quantum mechanics, and modeling spectral properties. With the increasing need for accurate prediction of atomic and molecular properties, interest is emerging in developing efficient computational methods to estimate the Gaunt factor, including the use of artificial intelligence techniques and modern algorithms[8].

1.2 Gaunt factor definition:

The Gaunt factor is a fundamental correction used in the calculation of radiative transfer rates, which improves the accuracy of these calculations by taking into account various physical effects [7, 9]. It takes into account the complex interaction between particle collisions and the processes associated with the emission or absorption of radiation. This factor plays a crucial role in understanding the behavior of matter when interacting with radiation, making it essential in many physical and astronomical applications [10].

1.3 Free-free Gaunt factor:

The cross-section for free-free absorption of radiation of frequency ν and energy $E_\nu = h\nu$ by electrons of initial energy E_a, final energy $E_b = E_a + h\nu$ and number

density N_e in a coulomb potential characterized by the central charge number Z:

$$\sigma_{ff}(\nu, a, b) = \sigma_{ff}^{k}(\nu, a; b) g_{ff}(\nu, a; b), \tag{1.1}$$

where

$$\sigma_{ff}^{k}(\nu, a, b) = (2/3m)^{3/2}(\pi Z^2 e^6/h\nu) N_e E_a^{-1/2} \nu^{-3}.$$

Is the Kramers' semi-classical cross-section, and $g_{ff}(\nu, a; b)$ is the quantum mechanical correction factor or Gaunt factor. It has been shown that the Gaunt factor may be written in terms of a partial wave expansion

$$g_{ff}(\nu, a; b) = (\sqrt{3/4})(\pi/Z^2)(h\nu/e)^4 \sum_{l_a m_a} \sum_{l_b m_b} | < a|r|b > |^2, \tag{1.2}$$

where l_a, m_a and l_b, m_b are the degenerate substate orbital and azimuthal angular momentum quantum numbers associated with the initial and final energy levels E_a and E_b, with $E_b = E_a + E_\nu$, and the dipole matrix element is defined with respect to initial and final free wave functions normalized on the energy scale. It is at once apparent that the expression for the Gaunt factor is symmetric with respect to the interchange of the initial and final levels. For a central potential the summation over the quantum numbers m_a and m_b can be immediately carried out to give:

$$g_{ff}(\nu, a; b) = (\sqrt{3/4})(\pi/Z^2)(h\nu/e)^4 \sum_{l_a} \sum_{l_b} \max(l_a, l_b)\delta(l_a - l_b, \pm 1)| < a|r|b > |^2 \tag{1.3}$$

For coulomb wave functions, the summation over the quantum numbers l_a and l_b can be expressed in terms of complex hyper-geometric functions [11] given by:

$$g_{ff}(\nu, a; b) = (2\sqrt{3})(\pi\eta_a\eta_b)[(\eta_a^2 + \eta_b^2 + 2\eta_a^2\eta_b^2) I_0 - 2\eta_a\eta_b\sqrt{1+\eta_a^2}\sqrt{1+\eta_b^2} I_1] I_0,$$

in which

$$I_1 = (1/4)[4k_a k_b/(k_a - k_b)^2]^{l+1} \exp[\pi|\eta_a - \eta_b|/2] \times [|\blacksquare(l+1+i\eta_a)\blacksquare(L+1+l\eta_b)|/\blacksquare(2l+2] G_l.$$

With the real function G_l given by:

$$G_l = |(k_b - k_a)/(k_b + k_a)|^{i(\eta_a+\eta_b)} \times_2 F_1[l+1-i\eta_b, l+1-i\eta_a; 2l+2; -4k_a k_b/(k_a - k_b)^2],$$

where the η and k are related to the energy E by $\eta = Z/(E/R_y)^{1/2}$ and the subscripts

a and b on all parameters denote their initial and final values. The calculation of G_I is calculated following the procedure of [11], more details are presented in [12].

1.4 Approximate formulae

Several approximate analytical formulae for the free-free Gaunt factor are available:

- $\eta_a >> 1$, $\eta_b < \eta_a$

$$g_{ff}(\nu, a; b) = 1 + A(1 + x^2)/y^{2/3} - B(1 + 484x^2/15 + x^4)/y^{4/3}, \tag{1.4}$$

  where

$$x = \eta_b/\eta_a$$

$$y(1 - x^2)\eta_b$$

$$A = \blacksquare(1/3)/\blacksquare(2/3)/12^{1/3}/5 = 1.72826... \times 10^{-1}$$

$$B = 3\blacksquare(2/3)/\blacksquare(1/3)/12^{1/3}/70 = 4.95957... \times 10^{-2}$$

  Valid when the initial electron energy is relatively high compared to the final energy, or when there is a large in electron energy during the process.

- $\eta_b << 1$,

$$g_{ff}(\nu, a; b) = 4\sqrt{3}\eta_b/\left[1 - \exp\left(-2\pi\eta_a\right)\right]. \tag{1.5}$$

  Appropriate when the final electron energy is very low.

- $\eta_b \approx \eta_a$

$$g_{ff}(\nu, a; b) = (\sqrt{3}/\pi)(\eta_a/\eta_b)\left[1 - \exp\left(-2\pi\eta_b\right)\right] / \left[1 - \exp\left(-2\pi\eta_a\right)\right] \ln\left[2\eta_b/(\eta_a - \eta_b)\right]. \tag{1.6}$$

  Which for $\eta_a << 1$ gives

$$g_{ff}(\nu, a; b) = (\sqrt{3}/\pi) \ln\left[(k_a + k_b)/(k_b - k_a)\right] \approx (\sqrt{3}/\pi) \ln(4E_a/E_\nu)$$

. Describe situations where the change in electron energy is relatively small.

1.5 The importance of Gaunt factor

The Gaunt factor is a dimensionless correction factor that appears in quantum mechanical formulations of atomic and radiative processes. It modifies classical

formulas for emission and absorption to account for quantum effects, particularly in free-free and bound-free transitions in plasmas. The mathematical relations involving the Gaunt factor link it to emission coefficients, absorption coefficients, and transition probabilities[13, 14, 15]. Below are the key relations and their connections to atomic parameters[16]:

- Primary Role: The Gaunt factor serves as a quantum mechanical correction to classical formulas describing radiative interactions (mainly free-free and bound-free transitions) in plasmas.
- Most Important Application: Its most significant use is in astrophysics, particularly in the modeling of stellar atmospheres and interpreting radiation spectra emitted by hot, ionized gases.
- Radiative Processes Affected:

  - Free-Free Emission (Bremsstrahlung): Occurs when an electron is deflected by an ion and emits a photon.

  - Bound-Free Transitions: Involves ionization or recombination of an electron with an atom or ion.
- Purpose of the Gaunt Factor:

  - Corrects classical cross-sections by incorporating quantum mechanical effects.

  - Enhances accuracy in calculating emission and absorption coefficients in high-temperature plasmas.

Atomic and Physical Parameters Determined After Calculating the Gaunt Factor:

- Radiative Transition Rates: The probability per unit time of a photon being emitted or absorbed during a transition.
- Emission Coefficient (ϵ_ν): The power emitted per unit volume, frequency, and solid angle — essential for spectral intensity modeling.
- Absorption Coefficient (κ_ν): Describes how much radiation is absorbed per unit distance in a medium.
- Opacity: A measure of the medium's transparency to radiation; critical for understanding radiative energy transport in stars.

- Plasma Diagnostics Parameters:

  - Electron density and temperature.

  - Ionization states.

  - Radiation losses in fusion or astrophysical plasmas.

1.6 Conclusion

The Gaunt factor is of great importance in characterization of fine quantum interactions. The challenges associated with accurately calculating it have prompted the search for alternative, more efficient methods, such as artificial intelligence-based predictive models.By deepening our understanding of this factor and developing our tools for calculating it, we open new horizons for scientific research in atomic physics and quantum chemistry[17].

Chapter 2

Machine learning

2.1 Introduction :

Machine learning involves learning, reasoning, and making decisions based on data. It works by developing computer programs that analyze data, extract valuable insights, predict unknown properties, and suggest actions or decisions. The key distinction of machine learning is that automation programs improve their performance by learning from data [18].

This means that general-purpose programs are adapted to specific applications by adjusting their parameters based on observed data, known as training data. Essentially, machine learning can be viewed as a form of programming by example. One of its greatest advantages is its versatility; machine learning methods can be applied across various domains for practical use [19].

The concept of a generic computer program corresponds to a mathematical model Of the data. Machine learning methods are defined using mathematical principles, which describe relationships between different quantities or variables representing observed data and the desired outputs. A mathematical model provides a compact, precise representation of data, capturing key properties of the studied phenomenon [20].

The choice of model depends on the available data and the expertise of the machine learning engineer. When implemented in practice, mathematical models are translated into code that runs on a computer. However, to fully understand the behavior of a machine learning program, it is crucial to grasp the underlying mathematical concepts [21].

There are many predictive techniques in machine learning; herein, we will use support vector regression.

2.2 Types of machine learning

There are three types of machine learning:

2.2.1 Supervised learning

Supervised learning is a type of machine learning where an algorithm learns from labeled data [22]. This means the training data includes both the input features and the corresponding correct outputs. The algorithm's goal is to learn a mapping function that can predict the output for new, unseen inputs [23, 22].

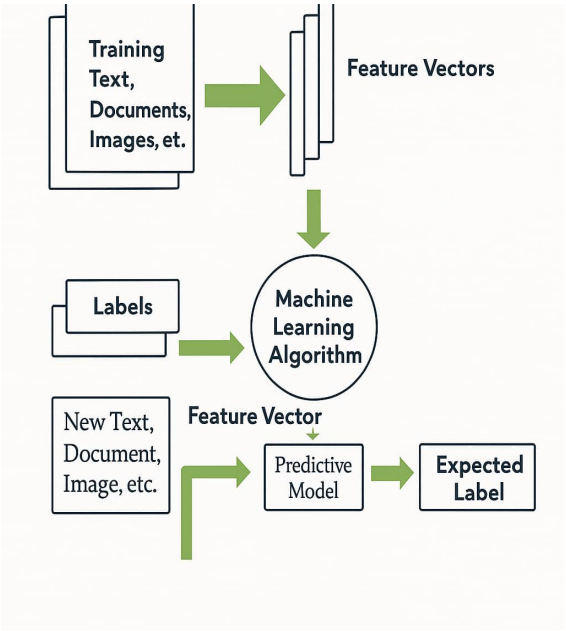

Figure 2.1: Supervised learning [1].

The figure(2.1) shows how a machine learning model is trained using labeled data and then used to predict labels for new data.

2.2.2 Unsupervised learning

Unsupervised learning is a type of machine learning where the algorithm learns patterns from unlabeled data [24]. Unsupervised learning discovers hidden patterns in data without labels. Since the data isn't categorized, the algorithm's output isn't judged for accuracy in the same way as supervised or reinforcement learning [22].

The figure(2.2) illustrates how an unsupervised learning model discovers patterns in unlabeled data.

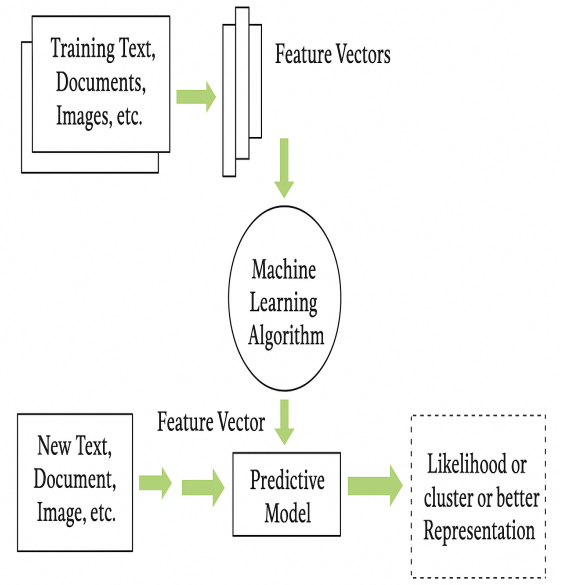

Figure 2.2: Unsupervised learning [1].

2.2.3 Reinforcement learning

Reinforcement learning is an ML technique that considers how intelligent agents should behave in a dynamic environment in order to maximize a predefined reward signal [24]. The agent receives feedback in the form of rewards and punishments, which it uses to guide its search for an optimal policy within the problem space [22].

2.3 Supervised Learning Techniques

2.3.1 Regression

Regression involves mapping input data to a numerical value.Given an input $X_i \in \mathbb{R}$ (representing a d-dimensional feature vector) and a continuous output space $y \subset \mathbb{R}$, the goal is to create a function f . $\mathbb{R} \rightarrow \mathbb{R}^{\times}$ that accurately maps any input X_i to its corresponding value $y \in \mathbb{R}$. Examples of regression methods include Neural Networks, Support vector regression, Linear Regression, and polynomial Regression [25].

Linear regression

A simple linear regression model for n observations can be expressed as:

$$y_i = \beta_0 + \beta_1 X_i + e_i. \tag{2.1}$$

Simple linear regression models the relationship between a single predictor variable X and a response variable y as a straight line. The term "simple" denotes that there's only one predictor, and "linear" indicates that the model is linear in the coefficients β The most common method for estimating β_0 and β_1 is the method of least squares. This method finds the line that minimizes the sum of the squared differences between the observed y_i values and the values predicted by the regression line and minimizes the following expression [26]:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 X_i)^2. \tag{2.2}$$

And the solution is:

$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i + \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2}, \tag{2.3}$$

$$\beta_0 = y - \beta_1 x. \tag{2.4}$$

Where n is the number of data.

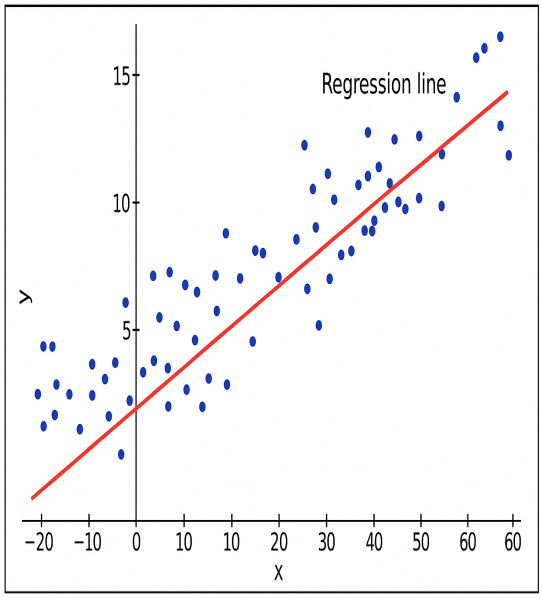

Figure 2.3: Linear regression [2]

The figure(2.3) shows a visual representation of linear regression, where the line

of best fit is plotted across sparse data.

Nonlinear regression

The fundamental concept of nonlinear regression is similar to that of linear regression, which aims to establish a relationship between a response variable Y and a set of predictor variables $X = (X_1, ..., X_k)^T$. What distinguishes nonlinear regression is that the prediction equation depends nonlinearly on at least one unknown parameter.

Linear regression is commonly applied when the primary goal is to build a purely empirical model. In contrast, nonlinear regression is typically used when there are theoretical or physical reasons to expect a specific fundamental relationship between the response variable and the predictors [27].

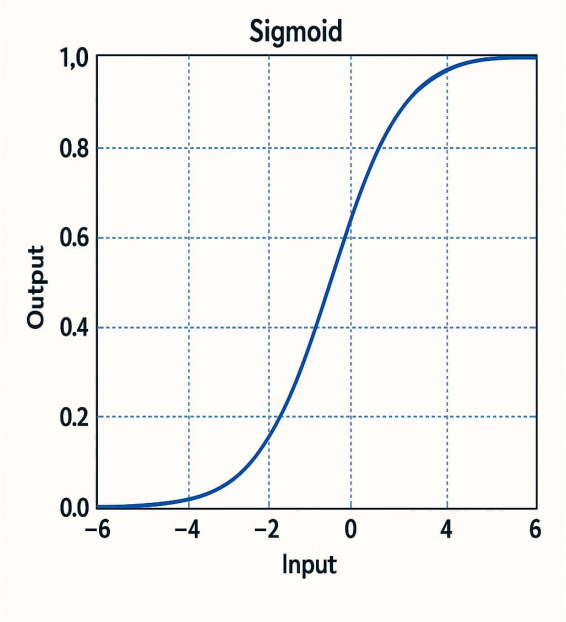

Figure 2.4: Nonlinear regression [3]

The figure(2.4) shows the visual representation of the nonlinear.

2.3.2 Classification

Classification is a supervised learning task where the goal is to predict categories [28]. Data mining is a key tool in machine learning, but people often make mistakes during analysis or when linking data. Machine learning can solve complex problems by improving system efficiency and machine design. It uses labeled (supervised) or unlabeled (unsupervised) data, with the letter aiming to uncover hidden patterns.

Many machine learning applications are supervised, following a typical classification architecture [29].

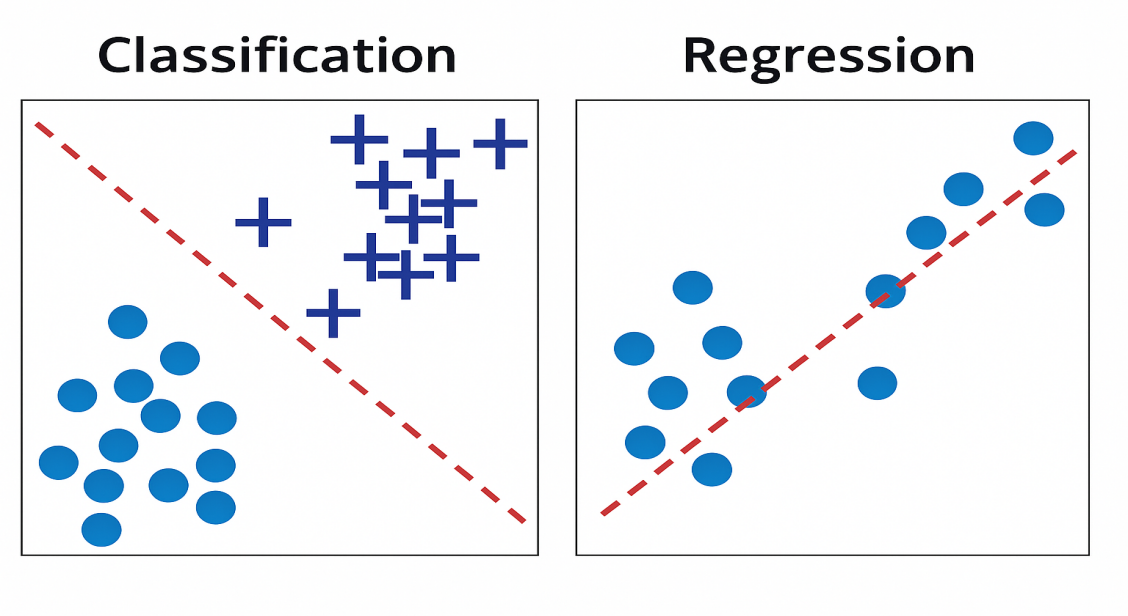

Figure 2.5: Difference between classification and regression [4]

The figure(2.5) shows that classification is concerned with assigning data to discrete classes, while regression aims to model the continuous relationship between variables.

2.4 Support Vector Regression for Prediction

The Support Vector Regression method aims to improve its ability to generalize to new data by carefully choosing the kernel function used. Selecting the appropriate kernel is crucial for the successful application of this technique. Support Vector Regression was initially derived from the theoretical concepts of Support Vector Machines.

Support Vector Machines can be adapted for regression tasks by employing the epsilon-insensitive loss function. This function helps evaluate the equality of the regression fit. When Support Vector Machines are used for regression, it is called Support Vector Regression.

The primary goal of Support Vector Regression is to identify a function $f(x)$ that deviates at most epsilon from the actual targeted values observed in the training data, While simultaneously ensuring the function is as smooth as possible. In simpler terms, errors within the range of epsilon are considered acceptable. In

Support Vector Regression, the training data points that influence the creation of the regression function are called support vectors. The function $f(x)$ is defined as:

$$f(x) = W^T \varphi(x) + b, \tag{2.5}$$

where $\varphi(x)$represents the result of mapping the input x into a higher-dimensional feature space, W is a weight vector, and b is the bias term. The parameters W and b are determined by minimizing a risk function [30]. The parameters W and b minimize the following risk function :

$$R = min\frac{1}{2}||W||^2 + (\sum_{i=1}^{l}(L_E(Y_i), f(x_i))) \tag{2.6}$$

Where L_E is the Loss function, SVR seeks a function $f(x_i)$ that exhibits the smallest deviation epsilon from the actual targeted values y_i for all training data points. Ideally, when epsilon is zero, perfect regression is achieved. However, a large epsilon value corresponds to a smaller influence of slack variables and leads to lower accuracy. Slack variables are introduced to address situations where it's impossible to maintain the margin defined by epsilon. The addition of this slack variable is to solve the problem of the infeasible margin limiter in the optimization problem [31].

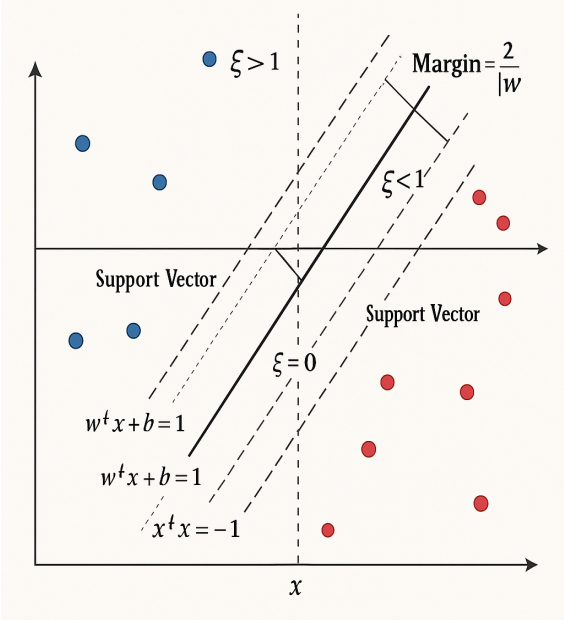

Figure 2.6: support vectors [5].

The figure(2.6) shows how the SVM works to find a decision boundary that

separates the classes while maximizing the margin, showing the role of support vectors and slack variables in the case of nonlinear separable data.

2.4.1 Application of SVR for prediction

Automated staging of an embryo :

SVR is used to develop a fast method for automatically determining the developmental age of a fruit fly embryo based on its segmentation gene expression patterns [32].

Geolocation problem in a mobile tracking scenario:

SVR is applied to estimate the location of a mobile user, with the addition of Kalman-Bucy filtering to smooth location estimates. This involves using radiolocation techniques and measurements like time of arrival (TOA),time difference of arrival(TDOA), angle of arrival(AOA), and signal strength(ss) [33].

Mobile Location Estimation:

SVR is used to obtain an initial estimate of the mobile location in a tracking scenario. A training database is created with measurements taken at known locations [34].

fMRI Data Analysis:

Spatio-temporal SVR is utilized for analyzing functional magnetic resonance imaging(fMRI) data.This approach helps in motion estimation, noise removal, and incorporating multi-run, multi-subject, and multi-task studies [35].

Blind Identification of SIMO Channels:

SVR is employed to solve the problem of blind identification of single-input multiple-output (SIMO) channels, which is common in communications, sonar, and seismic signal processing [36].

Oceanic Disasters Search and Rescue Operation:

SVR is used for system identification of a nonlinear black-box model in an ocean model.This helps predict the position of a target in distress, aiding search and rescue units [37].

Position Control of Ultrasonic Motor(USM)

SVR is used to develop a position control method for USMs, addressing the challenge of strong nonlinearity caused by friction. An SVR controller is combined with a PI controller for nonlinear input-output mapping [38].

Gene Selection for Continuous Phenotypes:

SVR is applied to select discriminative genes for continuous phenotypes(like the extent of programmed cell death) in microarray gene expression data analysis. This uses ordinal regression with multiple thresholds to define hyperplanes for ordinal scales [39].

2.5 Model Training and Evaluation

Model Training: One of the main challenges in machine learning is ensuring that a model performs well on new, previously unseen inputs, rather than just the data it was trained on. This capability, known as generalization, is crucial for the model's effectiveness [5].

Model Evaluation: To assess machine learning models, the dataset was divided into training and validation sets. This ensures that predictions are tested on data that was not used during model training. Various performance metrics can be used to evaluate forecasting models, including R-squared(R^2)and Root-mean-square error($RMSE$). The R^2 metric measures how well the forecasting model explains the variability in the actual outcomes. The $RMSE$ metric quantifies the differences between actual and predicted values, providing insight into the model's prediction accuracy.

$$R^2 = 1 - \frac{\sum_{i=1}^{M}(y_i - \hat{y}_i)2}{\sum_{i=1}^{M}(y_i - \bar{y})2} \tag{2.7}$$

$$RMSE = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(y_i - \hat{y}_i)2} \tag{2.8}$$

In these equations,M represents the total number of samples, y_i denotes the actual values, $\hat{y}_i$ refers to the predicted values, and $\bar{y}$ is the mean of the actual values [40].

2.5.1 Cross-validation and hyperparameter tuning

Cross-validation is a technique for selecting models and evaluating their performance by splitting data into training and testing sets. It estimates the error of

algorithms by training on one part and testing on another, ensuring independent validation. The most common method, k-fold CV, repeats this process k times, rotating test data to obtain multiple error estimates [41]. While cross-validation is used to build the final model, that final model itself isn't cross-validated. Instead, cross-validation gives an estimate of how well the final model will generalize to new data. Nested cross-validation separates the tuning of model settings from the evaluation, but simpler, standard cross-validation is often sufficient despite potentially overestimating performance [42].

2.5.2 Performance metrics

Root Mean square error

The RMSE, also called the root-mean-squared deviation, is a measure of the differences between the values predicted by a prediction model and the values actually observed, defined as:

$$RMSE = \sqrt{MSE} \tag{2.9}$$

and

$$MSE = \frac{1}{M}\sum_{i=0}^{M}(\hat{y}_i - y_i)^2 \tag{2.10}$$

The RMSE is a good measure of accuracy, but only to compare different predictions errors for a particular variable and not between variables, because it is scale-dependent [43].

Mean absolute error

The MAE is a quantity used to measure how close the estimated performance degradation trend $\hat{y}$ (or estimates) is to the actual performance degradation trend y (or actual responses), defined by:

$$MAE = \frac{1}{M}\sum_{i=0}^{M}|\hat{y}_i - y_i| \tag{2.11}$$

The MAE is also known as a scale-dependent accuracy measure and therefore cannot be used to make comparisons between series using different scales[[43]].

2.6 Unsupervised Learning Approaches

2.6.1 Clustering and dimensionality reduction

The traditional method for unsupervised word sense disambiguation involves automatically grouping instances of a word with multiple meanings based on their context. This is typically done by representing the context as a vector, similar to how documents are represented in information retrieval. These context vectors are then clustered using algorithms, aiming to group instances with the same word sense together [44].

Early approaches by Zernik and Schutze used this vector space model and explored different clustering methods. Zernik used hierarchical agglomerative clustering, while Schutze employed Bayesian and randomized algorithms, along with dimensionality reduction techniques. A major challenge with these unsupervised methods is the difficulty in matching the resulting clusters to predefined senses in a dictionary or inventory. Both Zernik and Schutze had to manually align the automatically generated clusters to known word senses after the clustering process. This manual step was necessary because the unsupervised clustering, based solely on context similarity, didn't inherently correspond to standard sense distinctions. They often generated multiple clusters per word sense and then manually labeled them [45].

2.6.2 Applications of unsupervised learning

Internet Traffic classification:

Internet traffic classification is essential for service providers to understand network characteristics like quality of service, user behavior, and security. Traditional port-basses classification is outdated due to dynamic port negotiation by malicious software. Modern methods use machine learning and clustering to classify packets by application, enabling traffic control, intrusion detection, and blocking of unwanted applications. Feature selection is crucial for accurate classification, with methods including filter, wrapper, and embedded approaches. Challenges remain in handling large data and imbalanced classes,leading to research in areas like ensemble feature selection and information-theoretic approaches [46].

Anomaly/Intrusion detection:

Are crucial for network security. Traditional signature-based systems, while precise in identifying known attacks, struggle with novel threats. Modem ADS leverages

unsupervised machine learning due to its ability to detect unknown behavior by establishing a baseline of normal activity and flagging deviations. Clustering algorithms like density-based, fuzzy rough C-means are employed to group similar network activity, with outliers considered anomalous. While effective against sophisticated attacks and insider threats, ADS has drawbacks: training and maintaining user profiles is challenging, false alarm rates can be high, and malicious users might manipulate the system over time [47].

Network operations, optimizations, and analytics:

Network management encompasses the operations of setting up, monitoring, and maintaining a computer network to ensure its basic functions are working correctly. The goal of network management and monitoring systems is to guarantee these functions are fulfilled, and analyze traffic patterns and suggest solutions to improve efficiency [48].

Dimensionality reduction and visualization:

Dimensionality reduction is the process of reducing the number of variables or features that describe the data. Imagine you have a very high-dimensional image, containing fine details that the naked eye cannot perceive. Dimensionality reduction helps remove this extra detail while preserving the key visual elements that define the image. Dimensional reduction techniques: Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), t-SNE, Autoencoder [49].

Other applications: Clustering customers data and market segmentation, learning rule associations, image segmentation, and gene clustering [50].

2.7 Challenges and Limitations of Machine Learning

2.7.1 Overfitting

Overfitting occurs when a machine learning model learns the training data too thoroughly, including its noise, leading to poor performance on new, unseen data. Essentially, the model memorizes the training data instead of generalizing. This is more common with complex models. Techniques like dropout can help mitigate overfitting [51].

2.7.2 Underfitting

Underfitting happens when a machine learning model is too simple and doesn't capture the underlying patterns in the data. This can be due to using too few predictors or unrepresentative training data. An underfit model performs poorly on both training and new data, as it struggles to generalize [51].

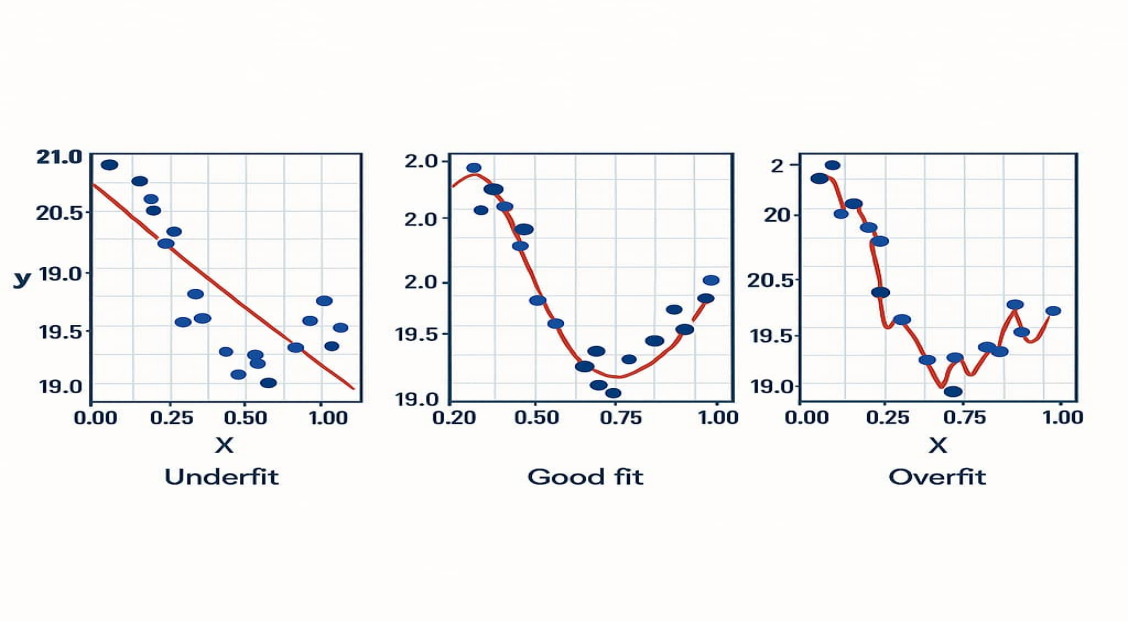

Figure 2.7: The difference between overfitting, underfitting [6].

The figure(2.7) shows how a model can be too simple (underfitting),too good (good fit), or too complex (overfitting) relative to the data.

2.7.3 Model complexity

The process of applying a machine learning algorithm involves experimentation to determine its suitability for the data and to discover ways to enhance its performance. In machine learning, we often have multiple potential models to choose from. The process of selecting the best model is known as model selection. The goal of a machine learning algorithm is to accurately predict outcomes for new, unseen data-not just the data it was trained on. This ability to make accurate predictions on new data is called generalization. Optimal generalization occurs when the complexity of the chosen model matches the complexity of the underlying within the data [52].

If a model is too simple to capture the underlying patterns in the data, it leads to underfitting.for example, in a straight line is used to fit data that likely follows

a higher-order polynomial. As we increase the model's complexity, the error on the training data decreases, resulting in a better fit. Here "error" refers to the sum of the squared distances between the data points and the model's predictions. A model that is excessively complex, especially when trained on limited data, can lead to overfitting. Imagine trying to fit a high-degree polynomial to data that actually follows a simpler pattern. While such a complex model might perfectly memorize the training data, achieving a very low training error, it will likely fail to generalize to new, unseen data. The risk of overfitting decreases with larger datasets. With enough data, a complex model can, after training, approximate a simpler one, achieving a more appropriate fit. However in practice, we often cannot guarantee an abundance of data.Furthermore visually comparing model complexity to the underlying data distribution is usually not possible. Therefore, we need alternative methods to evaluate model performance [23].

2.8 Conclusion

In this chapter we presented an overview of the importance and the utility of machine learning. Machine learning strategies introduce innovative approaches to leverage computational power and information across various scientific domains. They enable the analysis of large datasets within a relatively short time, a feat unattainable through manual efforts. This capability allows scientists to develop new experimental procedures and focus their efforts on the most promising questions within their field. However, computerized solutions are not a substitute for sound medical judgment. Like any other tool, machine learning techniques must be applied carefully to maximize their effectiveness. It is advisable to begin with simpler methods to assess problem complexity and gain deeper insights into algorithmic behavior. Additionally, exploring diverse algorithms and comparing their performance is crucial for achieving optimal results.

Chapter 3

Results and Discussion

3.1 Introduction

Recently, the use of Machine Learning (ML) models has became very widespread in different fields as a prediction tool. ML is finding a growing number of applications in physics, and making a significant impact on Particle physics accelerators, astrophysics, cosmology, materials science, condensed matter physics, and fluid dynamics plasma physics.

The key aim of the work is to contribute to the cal calculation of the free-free Gaunt factor by utilizing one of the available and inexpensive alternative methods, thereby avoiding lengthy mathematical computations and the need to reference databases of atomic and molecular properties repeatedly. The primary aim of this study is to demon strate the utility of Support Vector Regression algorithm in predicting the free Gaunt factor. Our goal is to construct a robust predictive model capable of accurately estimating the free Gaunt factor by training them on a comprehensive dataset comprising various astrophysical parameters. By harnessing the power of SVR.

3.2 Data Collection and Preprocessing

Data collection is the first step in any data driven project, where information is obtained from various sources to ensure the accuracy and comprehensiveness of the analysis. Data can be primary, collected directly from questionnaires, interviews, or observations, or secondary, extracted from databases, reports, or open sources such as Kaggle and UC Berkeley. Data can be structured, such as tables and databases unstructured. Data is collected manually or automatically using application programming interfaces or data extraction techniques [53].

Data preprocessing is the process of preparing data before analyzing it. It includes cleaning up missing and duplicate values, dealing with outliers, transforming data through normalization or standardization, and encoding categorical data to become numeric. It also includes selecting important features and splitting data into training and test sets to ensure model accuracy and improve analysis performance [54, 55].

3.3 Methodology

We aim in our work to establish a machine learning model that can calculate the free-free Gaunt factor g_{ff}using $log(u)$and $log(\gamma^2)$ as inputs. Knowing that:

$$u = \frac{h\nu}{k_B T_e}; \gamma^2 = \frac{Z^2 R_y}{k_B T_e}, \tag{3.1}$$

We collected the data of the free-free Gaunt factor from published scientific papers. The dataset consists of 819 data points, each described by three features: $log(u)$ ranging from -4 to 4, $log(\gamma^2)$ ranging from -8 to 8, and the targeted variable, the free-free Gaunt factor Figure (3.1).

$\log \gamma^2$	log u								
	−4.00	−3.00	−2.00	−1.00	0.00	1.00	2.00	3.00	4.00
−8.00	5.528E+00	4.259E+00	3.002E+00	1.806E+00	8.424E-01	3.035E-01	9.801E-02	3.107E-02	9.826E-03
−7.80	5.528E+00	4.259E+00	3.002E+00	1.806E+00	8.425E-01	3.035E-01	9.802E-02	3.107E-02	9.827E-03
−7.60	5.528E+00	4.259E+00	3.002E+00	1.806E+00	8.425E-01	3.035E-01	9.803E-02	3.107E-02	9.828E-03
−7.40	5.528E+00	4.259E+00	3.002E+00	1.807E+00	8.426E-01	3.035E-01	9.804E-02	3.108E-02	9.830E-03
−7.20	5.528E+00	4.259E+00	3.002E+00	1.807E+00	8.426E-01	3.036E-01	9.806E-02	3.108E-02	9.831E-03
−7.00	5.528E+00	4.259E+00	3.002E+00	1.807E+00	8.427E-01	3.036E-01	9.808E-02	3.109E-02	9.834E-03
−6.80	5.528E+00	4.259E+00	3.002E+00	1.807E+00	8.428E-01	3.037E-01	9.810E-02	3.110E-02	9.836E-03
−6.60	5.528E+00	4.259E+00	3.002E+00	1.807E+00	8.429E-01	3.038E-01	9.814E-02	3.111E-02	9.840E-03
−6.40	5.528E+00	4.259E+00	3.002E+00	1.807E+00	8.431E-01	3.039E-01	9.818E-02	3.112E-02	9.845E-03
−6.20	5.528E+00	4.260E+00	3.002E+00	1.807E+00	8.433E-01	3.040E-01	9.823E-02	3.114E-02	9.850E-03
−6.00	5.528E+00	4.260E+00	3.002E+00	1.807E+00	8.436E-01	3.042E-01	9.829E-02	3.116E-02	9.857E-03
−5.80	5.528E+00	4.260E+00	3.002E+00	1.808E+00	8.439E-01	3.044E-01	9.838E-02	3.119E-02	9.866E-03

Figure 3.1: Temperature-averaged free-free Gaunt factor vs. γ^2 for different u .

We integrate in Table (3.1) the computation of descriptive statistics to summarize our dataset. The table includes the count, mean, standard deviation,minimum, and maximum values.

Additionally, the distribution of the free-free Gaunt factor g_{ff} using Kernel Density Estimation (KDE) is illustrated in Figure (3.2). The distribution appears to be skewed toward one. The majority of the values fall between zero and 2, while the remaining values are sparsely distributed between 2 and 6.

	Count	mean	std	min	25%	50%	75%	max
$log(u)$	819.00	1.00	5.24	-8.00	-3.60	1.00	5.60	10.00
$log(\gamma)$	819.00	0.00	2.58	-4.00	-2.00	0.00	2.00	4.00
g	819.00	1.51	1.40	0.00	0.85	1.03	1.81	5.52

Table 3.1: Statistical summary of the dataset

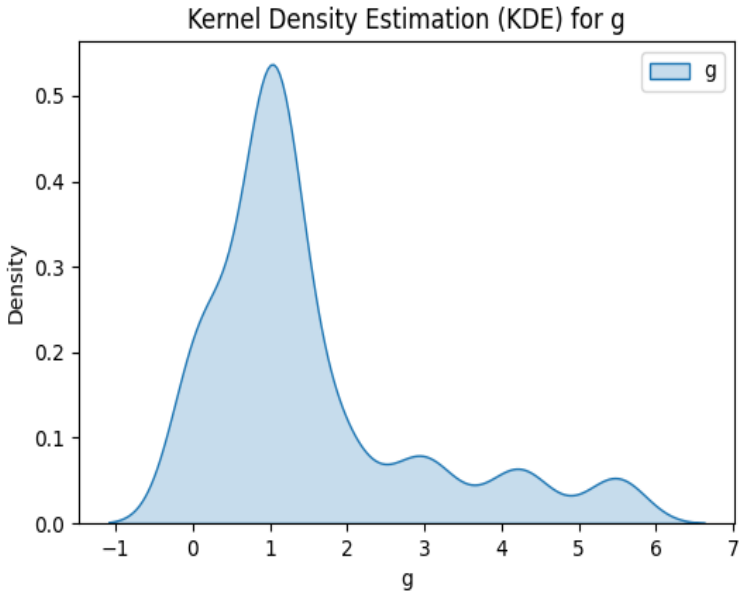

Figure 3.2: Dataset free-free Gaunt factor distribution

The figure(3.2) displays an estimated probability distribution for the variable 'g' using the kernel density estimation method.

3.4 Kernel functions:

The kernel function $K(x, y)$ takes two inputs (vectors) x and y computes an inner product in a high-dimensional space without the need to explicitly transform the data into that space. in other words, it provides a way to efficiently compute the relationship between two points in a nonlinear space.

Kernel function is a mathematical function used in machine learning algorithms, to determine the similarity between two points in an input space. A kernel function allows calculations to be performed in a high-dimensional space without having to perform an actual transformation for each point, reducing computational complexity.

Kernel functions allow SVR to handle nonlinear relationships in the data by implicitly mapping them to a higher-dimensional feature space. SVR enables effi-

cient handling of data that cannot be linearly separated in native space, Choosing the appropriate kernel function can significantly improve the accuracy of the SVR model.

3.4.1 Types of kernel functions :

Linear kernel:

The linear kernel is the simplest type of kernel function and is based on SVR multiplication used in the interior between points in the input space. It can be represented by the following equation:

$$K(x, x') = x.x' \tag{3.2}$$

Where x and x' are feature vectors.

A Linear kernel is used when the relationship between inputs and outputs is linear, that is, when the data can be represented by a straight line or a plane in higher dimensions. SVR Model equation with linear kernel of the form:

$$f(x) = w.x + b \tag{3.3}$$

Where w is a vector of weights, b is the constant term, x is the feature vector w and b are chosen so that the error between the predicted and actual values is minimized while adhering to a pre-specified ϵ-tube margin.

Polynomial kernel:

Polynomial kernel is a kernel function used to capture nonlinear relationships between variables by representing them in a higher-dimensional space using polynomials. It is defined by the following equation:

$$K(x, x') = (x.x' + c)^d \tag{3.4}$$

Where x and x' are the feature vectors, c is a constant used to control the effect of the linear term,d is the degree of the polynomial and the basic factor in determining the complexity of the polynomial model. We use a polynomial kernel in SVR when the relationship between variables is not linear but can be represented by a polynomial.

SVR Model equation in polynomial kernel of the form:

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) + b \tag{3.5}$$

Where α_i are the support coefficients,$K(x_i, x)$ is the output of the polynomial kernel function, b is the constant term.

Gaussian kernel or RBF kernel(Radial Basis Function):

RBF kernel is one of the most widely used kernel functions because it has a high ability to handle SVMs and SVR nonlinear data. This kernel is based on measuring the similarity between points using a function and is represented by the following equation:

$$K(x, x') = \exp(-\frac{||x - x'||^2}{2\sigma^2}) \tag{3.6}$$

Where x and x' are feature vectors, and σ is the control factor in the range influence between points,$||x - x'||^2$ is the square of the euclidean distance between points.

SVR Model with RBF kernel of the form:

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) + b \tag{3.7}$$

Where α_i are the support coefficients,$K(x_i, x)$ is the output of the RBF kernel function,b is the constant term.

sigmoid kernel:

It is a kernel function used in the sigmoid kernel to create a nonlinear model.SVM and SVR are based on the sigmoid function, which is often used in artificial neural networks. It is defined by the following equation:

$$K(x, x') = \tanh(\gamma(x.x') + c) \tag{3.8}$$

Where x and x' are the feature vectors,γ is a parameter that controls the effect between points,c is the constant helps in adjusting the kernel function, A sigmoid kernel is used when we want to simulate the nonlinear behavior of data. It can generate nonlinear decision boundaries, but in a different way than the RBF and

polynomial kernel. SVR Model equation in sigmoid kernel of the form:

$$f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x) + b \tag{3.9}$$

Where α_i are the support coefficients,$K(x_i, x)$ is the output of the sigmoid kernel function, b is the constant term.

3.5 Results and Analysis:

In this study, we employed a machine learning model for predicting the Gaunt factors SVR where SVR is effective in high-dimensional spaces. We utilized two datasets for our experiments. Each dataset was standardized to ensure comparability and improve model performance. The datasets were divided into training (80%) and testing (20%) sets, and we performed cross-validation to validate the model's robustness.

Hyperparameters for the model SVR were optimized using Optuna, a hyperparameter optimization framework. The key parameters tuned were the regularization parameter and the kernel coefficient.

The performance of the SVR was evaluated using k-fold cross-validation. The dataset is divided into k equally sized folds(in our case $k = 5$) and the model is then trained and tested k time, each time $(k - 1)$ of the folds as a training set and the remaining fold for testing the model. After network training, the regression accuracy metrics and statistical tests are determined to evaluate the net work's performance. The regression accuracy metrics are commonly used to quantify the differences between the actual and predicted values of the obtained model. The regression accuracy metrics and statistical tests include Root Mean Squared Error ($RMSE$) and Coefficient of de termination (R^2). In general, the optimal model is the one with the highest (R^2) and the lowest ($RMSE$). These metrics are defined as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{M}(y_i' - y_i)2}{\sum_{i=1}^{M}(y_i - \hat{y})2} \tag{3.10}$$

$$RMSE = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(y_i' - y_i)2} \tag{3.11}$$

Where y_i' represents the predicted values generated by the model, y_i' is the actual value, $\hat{y}$ is the average of the actual values and N stands for the total number of

samples. The model performance metrics are summarized in Table (3.2).

	Polynomial		RBF		Linear		Sigmoid	
KFold	R^2	**RMSE**	R^2	**RMSE**	R^2	**RMSE**	R^2	**RMSE**
Fold 1	0.9648	0.2638	0.9994	0.0333	0.9454	0.3288	0.6779	0.7983
Fold 2	0.9508	0.2941	0.9995	0.0307	0.9245	0.3644	0.6779	0.7154
Fold 3	0.9586	0.2849	0.9995	0.0324	0.9360	0.3542	0.6705	0.8036
Fold 4	0.9576	0.2763	0.9992	0.0378	0.9410	0.3259	0.6705	0.7059
Fold 5	0.9670	0.2799	0.9997	0.0270	0.9501	0.3440	0.6712	0.8835
Average	0.9598	0.302	0.9946	0.0322	0.9394	0.3435	0.6736	0.7599

Table 3.2: Predictive performance of the SVR model.

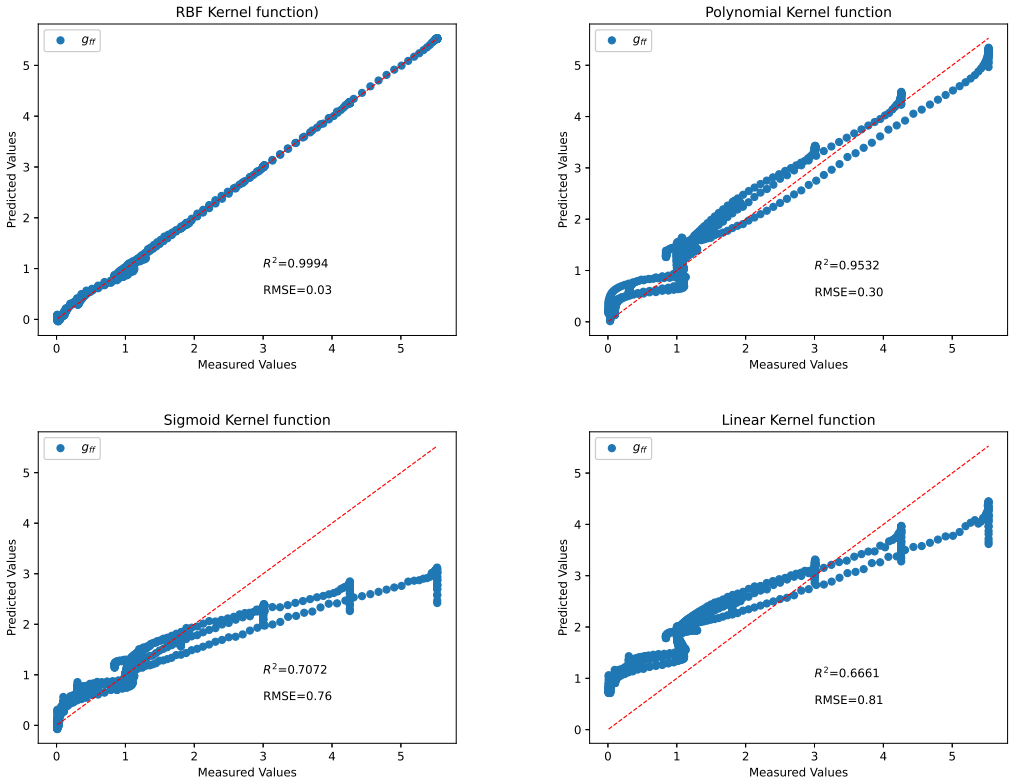

Figure 3.3: Plot of the measured versus predicted values and svr line(RBF, Poly, Sigmoid, Linear)

We evaluated the performance of SVR using four different kernel functions: **Radial Basis Function (RBF), polynomial, sigmoid**, and **linear**. The results were assessed by comparing the predicted values to the measured (true) values using standard metrics such as the coefficient of determination (R^2) and mean squared error (MSE). The corresponding plots in Figure (3.3) visualize the alignment between

predicted and measured values for each kernel, offering insight into each kernel's suitability for this regression task.

1. RBF Kernel

- Performance: $R^2 = 0.9994$, RMSE = 0.03
- The RBF kernel demonstrated outstanding performance, with an almost perfect fit to the measured values. The predicted points lie very closely along the ideal diagonal line, indicating that the model was able to capture the complex, possibly non-linear relationships within the data with high precision.
- The low MSE value further confirms that the prediction errors are minimal across the entire dataset.
- This result highlights the strength of the RBF kernel in handling non-linearity through its ability to assign influence based on radial distance in high-dimensional space. It effectively models localized variations and smooth transitions, making it especially suitable for accurately approximating the Gaunt factor.

2. Polynomial Kernel

- Performance: $R^2 = 0.9532$, RMSE = 0.3
- The polynomial kernel achieved good performance, though it did not match the precision of the RBF kernel. While the overall trend between predicted and measured values is still reasonably aligned, there is noticeable deviation, particularly at higher values of the Gaunt factor, where the model tends to show skewness and less consistent predictions.
- This suggests that while the polynomial kernel is capable of modeling non-linear relationships, it may overfit or underfit depending on the degree chosen and the distribution of data. It captures general trends but may struggle with localized fluctuations or subtle complexities in the data.

3. Sigmoid Kernel

- Performance: $R^2 = 0.6788$, RMSE = 0.80
- The sigmoid kernel exhibited relatively weak performance compared to RBF and polynomial. The scatter plot of predicted versus measured values shows a significant spread and a clear divergence from the ideal line, especially in mid- and high-value regions.

- The elevated MSE and lower R^2 indicate that the model fails to generalize well across the range of inputs, likely due to the sigmoid kernel's limited ability to model complex non-linear functions in this context.
- The sigmoid kernel behaves similarly to neural network activation functions, and while it may be suitable for binary classification tasks, it appears less effective for high-precision continuous regression in this specific problem.

4. Linear Kernel

- Performance: $R^2 = 0.7632$, RMSE = 0.68
- The linear kernel, like the sigmoid, showed limited predictive capability. Its performance is slightly better than the sigmoid in terms of R^2, but nearly identical in terms of MSE.
- This result is expected since the linear kernel is best suited for problems where the relationship between input and output is approximately linear — a condition clearly not met in our Gaunt factor dataset. As such, the model fails to capture the curvature and non-linearity present in the data, leading to systematic prediction errors.

5. Summary and Interpretation

The comparative analysis clearly demonstrates that the RBF kernel is the most effective for predicting the Gaunt factor, offering near-perfect accuracy and minimal error. Its flexibility in modeling non-linear patterns allows it to capture the intrinsic behavior of the data more effectively than other kernels. The polynomial kernel performs reasonably well but may introduce errors at the extremes. In contrast, both sigmoid and linear kernels fall short in this application due to their limited ability to model the non-linear nature of the underlying relationships.

These results underscore the importance of selecting an appropriate kernel function in SVR when dealing with physical parameters like the Gaunt factor, where accuracy and physical realism are critical. The superior performance of the RBF kernel makes it the most suitable choice for further development and integration into predictive models in plasma physics and astrophysics.

3.6 Conclusion

Our study demonstrates the efficacy of machine learning algorithms, specifically SVR, in accurately predicting the Gaunt factor. Leveraging a dataset comprising

819 Gaunt factor values from various sources. The model showed high accuracy in predicting the free-free Gaunt factor.

Our results clearly stand out the potential of machine learning algorithms in accurately estimating the free-free Gaunt factor, offering valuable insights for astrophysical research and exploration. Future work could explore additional machine learning models, larger and more diverse datasets, and potential applications to different types of plasma physics problems.

General conclusion

This work represents a step towards employing artificial intelligence and machine learning techniques in the service of theoretical physics by developing a predictive model based on the SVR algorithm to estimate the values of the Gaunt factor, which is one of the essential elements in complex quantum calculations. The results showed that using SVR can provide an efficient and accurate way to speed up calculations and reduce the computational costs associated with traditional methods without sacrificing the quality of the results.

Relying on a well-trained model to predict parameters of a mathematically complex nature opens the way for broader applications in multiple fields such as solid-state physics, theoretical chemistry, and materials science. This approach also provides a framework that can be generalized to other parameters with a similar structure and physical role.

In light of the encouraging results, this study recommends expanding the database used, experimenting with other machine learning algorithm's, and integrating hybrid models that combine traditional physics and machine learning, which may contribute to improving the accuracy of predictions and raising the efficiency of models in the future.

Bibliography

[1] Jafar Alzubi, Anand Nayyar, and Akshi Kumar. Machine learning from theory to algorithms: an overview. In *Journal of physics: conference series*, volume 1142, page 012012. IOP Publishing, 2018.

[2] unknown. Statistics - (univariate|simple|basic) linear regression, n.d. URL http://gerandico.com/wiki/data_mining/simple_regression. 27 April 2025.

[3] Marcel Caraciolo. Machine learning with python - logistic regression, Sunday, November 6, 2011. URL http://aimotion.blogspot.mk/2011/11/machine-learning-with-python-logistic.html. 27 April 2025.

[4] Yves Matanga Ngoma et al. *Analysis of Control Attainment in Endogenous Electroencephalogram Based Brain Computer Interfaces*. PhD thesis, Tshwane University of Technology, 2017.

[5] Salima BENBOUZID and Bilal KIR. *Prediction of some physical properties of metallic glasses using Machine Learning*. PhD thesis, University of Kasdi Merbah Ouargla.

[6] Solveig Badillo, Balazs Banfai, Fabian Birzele, Iakov I Davydov, Lucy Hutchinson, Tony Kam-Thong, Juliane Siebourg-Polster, Bernhard Steiert, and Jitao David Zhang. An introduction to machine learning. *Clinical pharmacology & therapeutics*, 107(4):871–885, 2020.

[7] PAM Van Hoof, RJR Williams, K Volk, Marios Chatzikos, Gary J Ferland, M Lykins, RL Porter, and Ye Wang. Accurate determination of the free–free gaunt factor–i. non-relativistic gaunt factors. *Monthly Notices of the Royal Astronomical Society*, 444(1):420–428, 2014.

[8] Brett I Dunlap. Generalized gaunt coefficients. *Physical Review A*, 66(3): 032502, 2002.

[9] Ralph S Sutherland. Accurate free—free gaunt factors for astrophysical plasmas. *Monthly Notices of the Royal Astronomical Society*, 300(2):321–330, 1998.

[10] DE Zenkhri, A Benkrane, and MT Meftah. Total free-free gaunt factors prediction using machine learning models. *Europhysics Letters*, 147(5):54001, 2024.

[11] WJ Karzas and Richard Latter. Electron radiative transitions in a coulomb field. *Astrophysical Journal Supplement, vol. 6, p. 167*, 6:167, 1961.

[12] TR Carson. Coulomb free-free gaunt factors. *Astronomy and Astrophysics (ISSN 0004-6361), vol. 189, no. 1-2, Jan. 1988, p. 319-324.*, 189:319–324, 1988.

[13] Christian Janicki. A computer program for the free-free and bound-free gaunt factors of rydberg systems. *Computer physics communications*, 60(3):281–296, 1990.

[14] Miguel A De Avillez and Dieter Breitschwerdt. Temperature-averaged and total free-free gaunt factors for κ and maxwellian distributions of electrons. *Astronomy & Astrophysics*, 580:A124, 2015.

[15] J Davis. Effective gaunt factors for electron impact excitation of multiply-charged nitrogen and oxygen ions. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 14(7):549–554, 1974.

[16] SM Younger and WL Wiese. An assessment of the effective gaunt factor approximation. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 22 (2):161–170, 1979.

[17] Narendra Nath Dutta. Trend of gaunt interaction contributions to the electric dipole polarizabilities of noble gas, alkaline-earth, and a few group-12 atoms. *Chemical Physics Letters*, 758:137911, 2020.

[18] Issam El Naqa and Martin J Murphy. What is machine learning? In *Machine learning in radiation oncology: theory and applications*, pages 3–11. Springer, 2015.

[19] Kiri Wagstaff. Machine learning that matters. *arXiv preprint arXiv:1206.4656*, 2012.

[20] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of

machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018.

[21] Andreas Lindholm, Niklas Wahlström, Fredrik Lindsten, and Thomas B Schön. *Machine learning: a first course for engineers and scientists*. Cambridge University Press, 2022.

[22] Diksha Sharma and Neeraj Kumar. A review on machine learning algorithms, tasks and applications. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, 6(10):2278–1323, 2017.

[23] Yalin Başţanlar and Mustafa Özuysal. Introduction to machine learning. *miRNomics: MicroRNA biology and computational analysis*, pages 105–128, 2013.

[24] Eleni Mamandra. Diabetes diagnonis using machine learning. Master's thesis, ■ανεπιστήμιο ■ειραιώς, 2022.

[25] Abderrahmane BENHADJIR. *Judd-Ofelt parameters: Bayesian inference and deep learning approach*. PhD thesis, 2021.

[26] Vladimir Nasteski. An overview of the supervised machine learning methods. *Horizons. b*, 4(51-62):56, 2017.

[27] Gordon K Smyth. Nonlinear regression. *Encyclopedia of environmetrics*, 3: 1405–1411, 2002.

[28] Elkhansa AIB. *Predicting of glass transition temperature of tellurite oxide glasses using Support vector regression*. PhD thesis, Université Kasdi-Merbah Ouargla.

[29] Iqbal Muhammad and Zhu Yan. Supervised machine learning approaches: A survey. *ICTACT Journal on Soft Computing*, 5(3), 2015.

[30] MF Rohmah, IKGD Putra, RS Hartati, and L Ardiantoro. Comparison four kernels of svr to predict consumer price index. In *Journal of Physics: Conference Series*, volume 1737, page 012018. IOP Publishing, 2021.

[31] Harsh S Dhiman, Dipankar Deb, and Josep M Guerrero. Hybrid machine intelligent svr variants for wind forecasting and ramp events. *Renewable and Sustainable Energy Reviews*, 108:369–379, 2019.

[32] Rebecca A Jones, Matthew J Renshaw, David J Barry, and James C Smith. Automated staging of zebrafish embryos using machine learning. *Wellcome Open Research*, 7:275, 2023.

[33] Fanzi Zeng, Shaoyuan Liu, Renfa Li, and Qingguang Zeng. Mobile tracking based on support vector regressors ensemble and game theory. *International Journal of Distributed Sensor Networks*, 10(3):403927, 2014.

[34] Satish R Jondhale, Vijay Mohan, Bharat Bhushan Sharma, Jaime Lloret, and Shashikant V Athawale. Support vector regression for mobile target localization in indoor environments. *Sensors*, 22(1):358, 2022.

[35] Xiaomu Song, Lawrence P Panych, and Nan-kuei Chen. Spatially regularized machine learning for task and resting-state fmri. *Journal of neuroscience methods*, 257:214–228, 2016.

[36] Steven VanVaerenbergh, Javier Via, and Ignacio Santamaria. Blind identification of simo wiener systems based on kernel canonical correlation analysis. *IEEE Transactions on Signal Processing*, 61(9):2219–2230, 2013.

[37] Xian-Rui Hou and Zao-Jian Zou. Svr-based parametric identification for parametric roll resonance of ships in longitudinal regular waves. In *International Conference on Offshore Mechanics and Arctic Engineering*, volume 49989, page V007T06A006. American Society of Mechanical Engineers, 2016.

[38] Hieu To Nguyen, Shogo Odomari, Tomohiro Yoshida, Tomonobu Senjyu, Atsushi Yona, and Vu Huu Thich. Digital position control strategy of traveling-wave ultrasonic motors. *Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije*, 55(3):246–255, 2014.

[39] Debasish Basak, Srimanta Pal, Dipak Chandra Patranabis, et al. Support vector regression. *Neural Information Processing-Letters and Reviews*, 11(10): 203–224, 2007.

[40] Amel A Alhussan, Mohamed S Gaafar, Mafawez Alharbi, Samir Y Marzouk, Sayer Alharbi, Hussain ElRashidy, Mai S Mabrouk, Hussah N AlEisa, and Nagwan Abdel Samee. Prediction of the judd–ofelt parameters of dy3+-doped lead borosilicate using artificial neural network. *Electronics*, 11(7):1045, 2022.

[41] Osval Antonio Montesinos López, Abelardo Montesinos López, and Jose Crossa. Overfitting, model tuning, and evaluation of prediction performance. In *Multivariate statistical machine learning methods for genomic prediction*, pages 109–139. Springer, 2022.

[42] Edson Duarte and Jacques Wainer. Empirical comparison of cross-validation and internal metrics for tuning svm hyperparameters. *Pattern Recognition Letters*, 88:6–11, 2017.

[43] Myeongsu Pecht, Michael G.; Kang. Prognostics and health management of electronics (fundamentals, machine learning, and the internet of things) ||. 10.1002/9781119515326, sep 2018. doi: 10.1002/9781119515326. URL libgen.li/file.php?md5=2063e79b7d1f3a28dcf5fa82e357eae6.

[44] Amruta Purandare and Ted Pedersen. Word sense discrimination by clustering contexts in vector and similarity spaces. In *Proceedings of the eighth conference on computational natural language learning (CoNLL-2004) at HLT-NAACL 2004*, pages 41–48, 2004.

[45] Nitin Indurkhya and Fred J Damerau. Handbook of natural language processing. *Computational Linguistics*, 37(2):395–397, 2011.

[46] Jingjing Zhao, Xuyang Jing, Zheng Yan, and Witold Pedrycz. Network traffic classification for data fusion: A survey. *Information Fusion*, 72:22–47, 2021.

[47] Usama Ahmed, Mohammad Nazir, Amna Sarwar, Tariq Ali, El-Hadi M Aggoune, Tariq Shahzad, and Muhammad Adnan Khan. Signature-based intrusion detection using machine learning and deep learning approaches empowered with fuzzy clustering. *Scientific Reports*, 15(1):1726, 2025.

[48] Alexander Clemm. *Network management fundamentals*. Cisco press, 2006.

[49] Muhammad Usama, Junaid Qadir, Aunn Raza, Hunain Arif, Kok-Lim Alvin Yau, Yehia Elkhatib, Amir Hussain, and Ala Al-Fuqaha. Unsupervised machine learning for networking: Techniques, applications and research challenges. *IEEE access*, 7:65579–65615, 2019.

[50] Hany Alashwal, Mohamed El Halaby, Jacob J Crouse, Areeg Abdalla, and Ahmed A Moustafa. The application of unsupervised clustering methods to alzheimer's disease. *Frontiers in computational neuroscience*, 13:31, 2019.

[51] Jeff Heaton. Ian goodfellow, yoshua bengio, and aaron courville: Deep learning: The mit press, 2016, 800 pp, isbn: 0262035618. *Genetic programming and evolvable machines*, 19(1):305–307, 2018.

[52] Isabelle Guyon. A practical guide to model selection. *Proc. Mach. Learn. Summer School Springer Text Stat*, pages 1–37, 2009.

[53] Kongmany Chaleunvong. Data collection techniques. *Training Course in Reproductive Health Research Vientine*, 2009.

[54] Sotiris B Kotsiantis, Ioannis D Zaharakis, and Panayiotis E Pintelas. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26:159–190, 2006.

[55] Sergio Ramírez-Gallego, Bartosz Krawczyk, Salvador García, Michał Woźniak, and Francisco Herrera. A survey on data preprocessing for data stream mining: Current status and future directions. *Neurocomputing*, 239:39–57, 2017.