

Applying Supervised Machine Learning for Radiation Dose Accumulation

UNIVERSITY OF TURKU
Department of Computing
Master of Science (Tech) Thesis
Health Technology
February 2025
Valtteri Puumalainen

Supervisors:
Jari Björne, PhD. (University of Turku)
Jussi Nieminen, M.Sc. (Teollisuuden Voima Oyj)
Eemeli Härmälä, M.Sc. (Teollisuuden Voima Oyj)

UNIVERSITY OF TURKU
Department of Computing

VALTTERI PUUMALAINEN: Applying Supervised Machine Learning for Radiation
Dose Accumulation

Master of Science (Tech) Thesis, 77 p.
Health Technology
February 2025

Predicting radiation doses in nuclear power plants is a challenging problem for maintaining the radiation safety of workers whilst ensuring that high exposures do not occur. These radiation doses can be predicted using different sensors, measurements or by manually reviewing the dose history of personnel. However, in this study, a previously unexplored visit-based machine learning approach for predicting radiation doses was developed. This approach utilises time relational data on personnel visits to the controlled area of OL1 and OL2 (Olkiluoto Unit 1 and 2) nuclear power plants, including radiation doses measured during these visits. This allows us to predict visits for different interval classes depending on the radiation dose received.

To provide a comprehensive foundation for machine learning modeling, we also examined the regulations governing current activities and analysed the nature of radiation exposure in nuclear power plant environments, including the origins and effects of radiation. Finally, we evaluated the prerequisites and considerations for deploying a comparable application in a production environment.

Through a combination of literature and experimental analysis, a basis for machine learning analysis was established, adopting five different models: 1) Random Forest, 2) Balanced Random Forest, 3) XGBoost, 4) LightGBM and 5) Easy Ensemble with AdaBoost. Among the models tested, LightGBM achieved the most promising results, however, its performance fell short of expectations due to the inherent imbalance and lack of descriptiveness in the dataset. While the models demonstrated an ability to learn from the data, this learning was insufficient to effectively distinguish between all class intervals. These limitations emphasise the value of integrating additional contextual information, such as the specific work tasks completed during visits, to enhance the dataset's descriptiveness and improve the model's performance.

By addressing these limitations, this study highlights the broader potential for data-driven modelling and further research. Specifically, we demonstrate that the descriptiveness and contextual relevance of data are as, if more, important as its quantity, as the mere existence or abundance of data does not guarantee its applicability to similar data-driven methods.

Keywords: machine learning, ml, radiation exposure, occupational exposure, radiation dose, radiation, as low as reasonably achievable (ALARA), nuclear power plant, npp, nuclear energy, crisp-dm

Acknowledgements

I would like to begin by thanking my supervisors, Jari Björne, Jussi Nieminen and Eemeli Härmälä, for their good and constructive approach to my work. I felt that I was contributing to something meaningful, both for my own career and for the company that made my research possible, Teollisuuden Voima Oyj. In addition, I would also like to express my gratitude to everyone else who has helped me with my work, both with the databases and with the data, tools and systems. Your support has been invaluable and has greatly accelerated my progress over the past year.

Furthermore, I am grateful to my previous employers, and my current employer, Teollisuuden Voima Oyj, for enabling me to be here today. The opportunity to learn and grow, which is of great significance at the start of one's career, has been truly invaluable. I also extend my sincere thanks to everyone at Turun Yliopisto (University of Turku) for teaching and supporting me throughout my studies. This is an excellent place to continue!

Valtteri Puumalainen

January, 2025

Contents

1	Introduction	1
2	Research process	17
3	Radiation and machine learning	20
3.1	Legislative approach to dose regulation	20
3.2	Nuclear power plant setting	22
3.2.1	Site visit formation	23
3.2.2	Radiation phenomena and dosage	24
3.2.3	Biological effects of radiation	29
3.3	Related studies and machine learning	33
3.3.1	Present-day applications	34
3.3.2	Overview of machine learning	36
4	Machine learning approach	45
4.1	Data	45
4.2	Applying machine learning	56
4.3	Evaluation	58
5	Results	63
5.1	Evaluation analysis	71
5.2	Operational use	72

6 Conclusions	76
References	78

List of acronyms

$H^*(d)$ Ambient Dose Equivalent

$H_p(d)$ Personal Dose Equivalent

w_R Radiation Weighting Factor

w_T Tissue Weighting Factor

A Mass Number

AdaBoost Adaptive Boosting

ALARA As Low As Reasonably Achievable

ANN Artificial Neural Network

AUC Area Under the Curve

Bq Becquerel

BRF Balanced Random Forest

BWR Boiling Water Reactor

CRISP-DM CRoss-Industry Standard Process for Data Mining

D-Tree Decision Tree

EE Easy Ensemble

EFB Exclusive Feature Bundling

GBDT Gradient Boosting Decision Trees

GBN Gaussian Bayesian Network

GDPR General Data Protection Regulation

GOSS Gradient-Based One-Side Sampling

GP Gaussian Process

Gy Gray

ICRP International Commission on Radiological Protection

IQR Interquartile Range

IXRPC International X-Ray and Radium Protection Committee or Commission

LDR Low-dose radiation

LightGBM Light Gradient Boosting Machine

LLN Law of Large Numbers

LNT Linear No-Threshold

LWR Light Water Reactor

MAE Mean Absolute Error

man-Sv man-Sievert

MAPE Mean Absolute Percentage Error

ML Machine Learning

NLL Negative Log-Likelihood

NPP Nuclear Power Plant

OOB Out-of-bag

OvR One-vs-the-Rest

PR Precision-Recall

PTSD Post-traumatic stress disorder

RF Random Forest

RMSE Root Mean Squared Error

ROC AUC Area Under the Receiver-Operating Characteristic

ROC Receiver Operating Characteristic

RWP Radiation Work Permit

STUK Radiation and Nuclear Safety Authority

Sv Sievert

TLD Thermoluminescent Dosimeter

TVO Teollisuuden Voima Oyj

WDC Work Dose Code

XGBoost Extreme Gradient Boosting

Z Atomic Number

1 Introduction

Radiation is a phenomenon that has always occurred naturally on Earth in two distinguishable forms: *ionising* and *non-ionising*. The defining characteristic of ionising radiation is its ability to remove electrons from atoms or molecules, thereby producing ionisation, whereas non-ionising radiation does not have enough energy to achieve this [1]–[3]. These types of radiation occur in nature as so-called background radiation due to cosmic rays and terrestrial radiation [1]. Although the general understanding is that radiation is perceived as harmful, it is often employed in healthcare settings for a variety of diagnostic and therapeutic procedures, as well as in other applications, such as nuclear power production [4]. This is referred to as *man-made radiation* [1].

The focus of this study is Teollisuuden Voima Oyj (TVO), the operator of three nuclear power plants (NPPs) in Olkiluoto, Finland. These NPPs account for one third of Finland’s electricity production, the first of which has been in operation since 1979 and the latest since 2023 [5]. To operate these NPPs, TVO must comply with the laws, decrees, decisions and regulations laid down in Finnish legislation, which are affected by the decisions of the Council of Europe [6]. The Radiation and Nuclear Safety Authority (STUK) oversees nuclear safety and radiation monitoring in Finland, and therefore sets regulations that need to be followed [6], [7].

This study focuses on man-made ionising radiation from fission reactions, resulting in fission products, to which personnel are exposed in the nuclear power plant

environments. This exposure may be referred to as occupational exposure [8]. TVO has a legal obligation to maintain and manage radiation dose data from their NPPs in order to monitor personnel's exposure [6]. The objective of this study is to utilise this data to predict the radiation dose of personnel during site visits using machine learning. At present, this prediction is carried out manually based on past experience and work history. This study therefore seeks to assess whether such prediction can be achieved through the application of machine learning, which can be defined as the teaching of an algorithm with historical data, thereby enabling it to adapt and make accurate predictions based on past experiences [9], [10]. More precisely, this problem will be approached through classification, which means that the radiation dose is predicted with a finite number of possible outcomes, even if the dose is a continuous value, but this problem is solved by dividing the value of the dose into interval classes [11].

The current manual method for predicting radiation doses achieves a Mean Absolute Error (MAE) of 111.9 man-mSv and a Mean Absolute Percentage Error (MAPE) of 18.76% for annual maintenance cycles completed between 2012 and 2023. This means that the average difference between the predicted and actual radiation doses is 111.9 man-mSv and the percentage error relative to the actual dose is 18.76% [10]. In the light of these metrics, the term man-Sievert (man-Sv) is used to describe the collective dose received by a group of people. It represents the sum of the individual doses within some population.

The data used in this study for the machine learning modelling was collected from the same period as the initial (MAE and MAPE) benchmarks, spanning from 2012 to 2023. The MAE and MAPE values are calculated from the total accumulated dose and dose predictions for both OL1 and OL2 at the annual maintenance level.

In this sense, the machine learning model is assumed to have at most the same MAE and MAPE values predicting the total dose. This study therefore sets as a

success criterion that the value obtained by machine learning prediction should be at least as accurate on the annual cycle level as that obtained by manual prediction.

These maintenances may include either a nuclear fuel change or maintenance that includes a nuclear fuel change. These take place alternately at OL1 and OL2. At the same time, the maintenance includes any necessary repair of malfunctions, possible modifications and, if necessary, preparation for the next year's maintenance. The necessity for annual maintenance represent a significant aspect of the operational life cycle of the plant units [5].

The following research questions are formed with the intention of modelling the effects of ionising radiation on humans in nuclear power plant environments. This will provide a basis for the processing of radiation dose data. The processing of the data must take into account the legal regulations and the operating environment. Finally, the most suitable machine learning model for the given problem will be identified, and the potential integration of such a model into a network information system solution will be explored. The study makes use of the CRISP-DM (CRoss-Industry Standard Process for Data Mining) process model.

- **Research question 1:** What is the impact of radiation on individuals in nuclear power plant environments?
- **Research question 2:** Which machine learning model is the most suitable for predicting occupational radiation exposure during a site visit?
- **Research question 3:** How can the machine learning model implemented in this research be integrated into real-world nuclear power plant operations?

The succeeding paragraphs present the general information on the subject of the study, which is essential for comprehending the overall context. This will enable the reader to understand the fundamental principles of nuclear power plant operation, radiation sources, and radiation monitoring activities that underpin nuclear and

radiation safety with the focus on *individual monitoring*, meaning the making and interpretation of measurements related to occupational exposure [8]. This will provide an overview that will guide the reader towards a more in-depth understanding of Chapter 3.

As previously stated, TVO operates three nuclear power plants, OL1, OL2 and OL3 (Olkiluoto Unit 1, 2 and 3). The first two plants are identical and each has a net electricity output of 890 MW (megawatt) [5]. In contrast to the preceding units, the third nuclear power plant, OL3, operates on a different principle, with the capacity to generate electricity with a net electricity output of 1600 MW [12]. This study will focus on OL1 and OL2. This choice is based on the fact that these sites have been in production use for a considerable length of time compared to OL3 and thus have well established work procedures and significantly more data available.

In principle, nuclear energy counts as thermal energy, and the OL1 and OL2 plants operate on this principle, using Boiling Water Reactors (BWRs), or more generally, Light Water Reactors (LWRs) [5]. BWRs operate by circulating water between the fuel rods within the reactor core, which results in the heating and vaporisation of the water [5]. In the case of TVO, the resulting steam is then directed through four main steam pipes to the high-pressure turbine, which directs the steam to be reheated with intermediate heater, before finally reaching the four low-pressure turbines that work in parallel [5]. These turbines drive the generator. The power of the reactor is controlled by control rods and main circulation pumps [5]. The reactors use uranium dioxide (UO_2) pellets as fuel, with the primary process being the neutron-induced fission of uranium-235 isotopes [3], [5]. This fission reaction, among others, releases ionising radiation, such as neutrons and gamma rays, and the necessary non-ionising radiation, heat [3], [5]. The BWR system of the OL1 and OL2 reactors is illustrated in Figure 1.1. A detailed overview of the operational aspects of the reactor is not within the scope of this study.

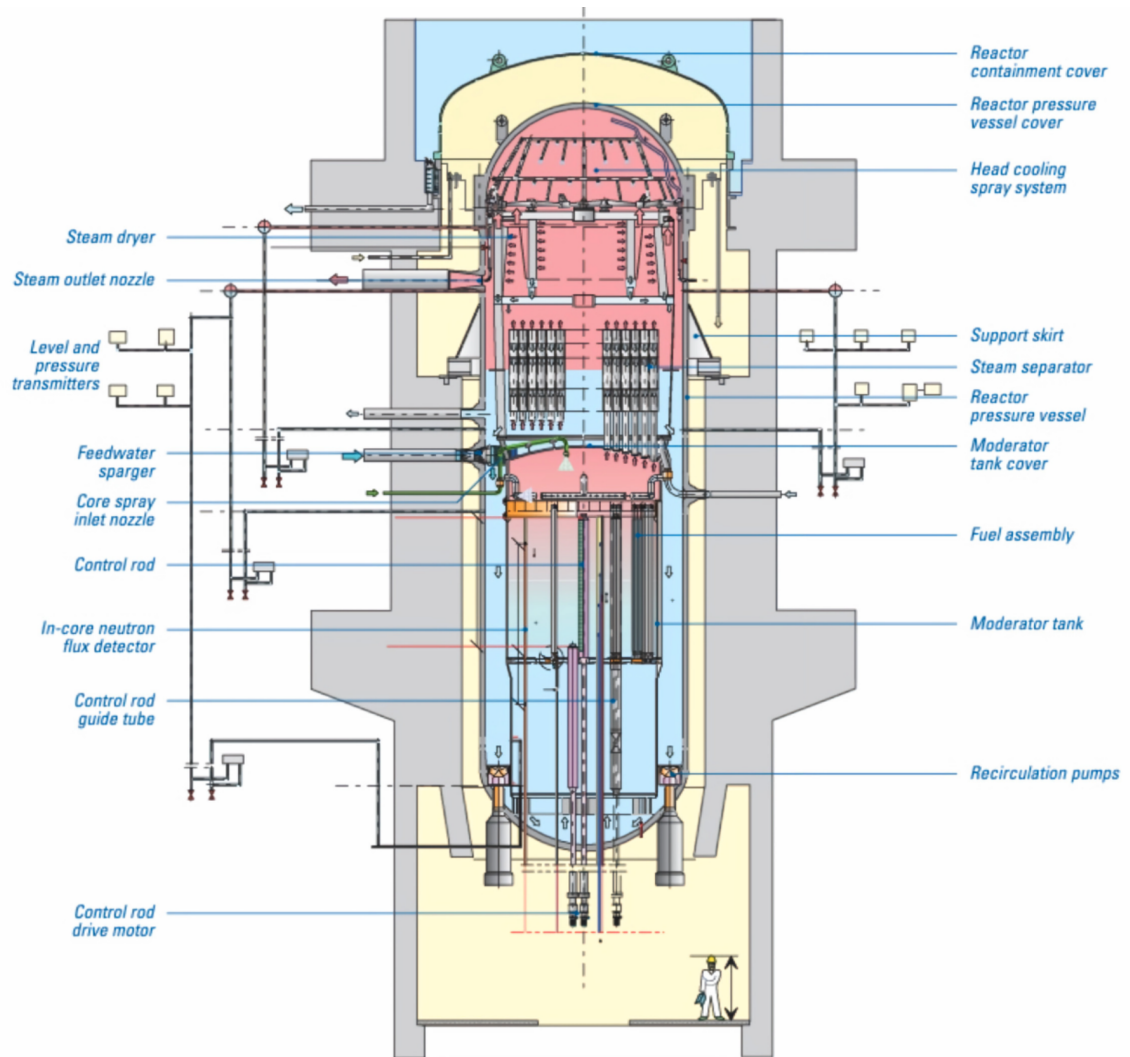


Figure 1.1: OL1 and OL2 BWR system overview [3], [13]

In consideration of the various radiation sources, the most significant is ionising radiation emitted from the reactor. However, this has been considered in the design of the NPPs and the fuel pellets utilised in the reactor already serve as the initial barrier against the dissemination of radiation [5]. The second layer of protection is the metallic shell of the fuel rods, which contain the fuel pellets [5]. The third layer of protection is the reactor pressure vessel, which is protected by a containment building that acts as fourth layer [5]. The final layer of protection is the reactor building [5]. These constitute a series of nested protective zones.

Despite the above-mentioned protective measures, radiation doses to personnel still occur, and the highest doses occur during the mandatory annual maintenance of plants, as this is when a lot of work is carried out near radiating objects. It should be noted that the plants emit more radiation during power operation, but in this case there is not as much work being carried out on the radiating systems as during maintenance, if any, so there is no increase in the dose of the personnel [5]. In this instance, the source of radiation is the contamination of water that comes into contact with the reactor. This water contains radioactive impurities due to the activation of elements like oxygen that transform into nitrogen-16 (^{16}N), which emits gamma radiation. These impurities accumulate in various systems, including on the surfaces of pipes. However, during a reactor shutdown when the maintenance is done, the radioactivity of the water decreases as short-lived isotopes like nitrogen-16 decay, reducing the overall radiation levels.

The radiation in question is produced as a consequence of radioactivity, whereby unstable atomic nuclei undergo a spontaneous change in state, emitting particles or electromagnetic radiation, in order to achieve stability [1], [2]. The aforementioned stability thus gives origin to ionising radiation, which can be classified into four distinct categories: 1) α (alpha) particles, 2) β (beta) particles, 3) γ (gamma) rays, 4) X-rays (X-rays) and 5) n (neutrons) [1]. Of these categories, gamma and X-ray radiation is discussed in detail in Chapter 3.2.2, as these are measured by the electronic dosimeters and forms the basis of the dose in the available data. It should be noted that TVO is also capable of measuring other types of radiation, but gamma radiation (and similarly, X-rays, due to their similar electromagnetic nature) is always measured in due of penetration capability and thus, the significance.

The basis of radioactivity lies in the atomic nucleus, which is composed of protons and neutrons, with protons exhibiting a charge magnitude comparable to that of electrons but opposite in sign, and neutrons remaining neutral as they bind these

protons [2]. Radionuclides with an unfavourable neutron-proton ratio undergo decay, thereby losing energy and becoming other nuclides or isotopes with different atomic numbers (Z) or mass numbers (A) [2]. The atomic number Z represents the number of protons present in an atom, while the mass number A is the sum of protons and neutrons [2]. A nucleus is considered stable when its neutron-to-proton ratio is approximately equal to one [2]. In general, nuclei with an atomic number greater than 83 are unstable, such as uranium ($Z=92$), which is often used fuel in NPPs [2], [5].

The process by which atomic nuclei seek equilibrium, resulting in the release of radiation, is invisible to the human eye and cannot be directly observed by humans [2]. However, radiation can be measured using dosimeters or with radiometers, which provide a quantifiable reading of the radiation levels. Before doing so, however, it must be understood that the amount of radioactivity is measured by the activity, which is the number of disintegrations of a radionuclide per unit time, which is measured in *Becquerel* (Bq) or reciprocal seconds (s^{-1}), where $1 \text{ Bq} = 1 \text{ s}^{-1}$ [3]. The reciprocal second refers to the frequency of events or decays in seconds in a radioactive material [3]. This can be further discussed in terms of absorbed dose, which is the amount of energy transferred by ionising radiation to a substance, i.e. a unit of mass [2], [3]. This is measured in *Gray* (Gy), where the unit is the joule per kilogram (J.kg^{-1}), where $1 \text{ Gy} = 1 \text{ J.kg}^{-1}$ represents the absorption of 1 joule of radiation energy per kilogram of matter [3]. Unlike absorbed dose, *Sievert* (Sv) takes into account the biological effects of radiation on different human tissues [2]. This is achieved by assigning different types of radiation a radiation weighting factor (w_R) [2]. The weighted values are then summed to calculate the *equivalent dose*, where the unit is J.kg^{-1} [2]. Table 1.1 shows these weighting factors and their respective continuous functions for neutron radiation as a function of energy.

Radiation type	Radiation weighting factor (w_R)
Photons (γ and X-rays)	1
Electrons (β)	1
Alpha particles (α)	20
Neutrons (n)	A continuous function of neutron energies:
$E_n < 1$ MeV	$w_R = \begin{cases} 2.5 + 18.2 \cdot e^{-[\ln(E_n)]^2/6} \\ 5.0 + 17.0 \cdot e^{-[\ln(2E_n)]^2/6} \\ 2.5 + 3.25 \cdot e^{-[\ln(0.04E_n)]^2/6} \end{cases}$
$1 \text{ MeV} \leq E_n \leq 50 \text{ MeV}$	
$E_n > 50 \text{ MeV}$	

Table 1.1: Radiation types and weight factors defined by ICRP (*International Commission on Radiological Protection*) in 2007 [2]

This brings us to the *effective dose*, which takes into account the radiation weighting factor as well as the varying sensitivity of different tissues and is again expressed in *Sieverts* [2], [3]. This is calculated by weighting the equivalent dose for each tissue or organ by the tissue weighting factor (w_T), whereby the sum of all tissue weighting factors for the body would be one, and then summing the result. Table 1.2 shows these weighting factors. Equivalent and effective doses cannot be calculated directly but are assessed using radiation dose quantities and modeled with computational phantoms, which are computer models of human anatomy composed of numerous voxels [2]. Each voxel is assigned a specific tissue type and organ identity based on gender. [2]. These phantoms simulate how radiation distributes in the body and are created using computational geometry or 3D imaging [2].

Tissue or organ	w_T	$\sum w_T$
Bone marrow, colon, lung, stomach, breast, and remainder tissues	0.12	0.72
Gonads	0.08	0.08
Bladder, esophagus, liver and thyroid	0.04	0.16
Bone surface, brain, salivary glands skin	0.01	0.04

Table 1.2: Tissue types and weight factors defined by ICRP in 2007. Remainder tissues include adrenals, nasal and oral passages, pharynx, larynx, gall bladder, heart, kidneys, lymphatic nodes, muscle, oral mucosa, pancreas, prostate, small intestine, spleen, thymus, uterus and cervix [2]

For operational calculations, there are definitions of *internal* and *external exposure*, of which this study only considers external exposure, as the data collected for this study only covers the external exposure of personnel. Internal exposure occurs when radioactive substances enter the body by ingestion, inhalation or absorption, for example, through wounds or breathing [2], [3]. External exposure, on the other hand, occurs when the body is exposed to external radiation from particles emitted by a radioactive source as we have previously discussed [2].

External exposure can be measured in ambient dose equivalent ($H^*(d)$) for area monitoring and personal dose equivalent ($H_p(d)$) for individual monitoring [2]. This study focuses on the *personal dose equivalent* as it is central to personal monitoring of occupational exposure. $H_p(d)$ is measured at a depth d of 10mm for the effective dose or 0.07mm for the skin dose, i.e. $H_p(10)$ is suitable for measuring the deeper penetrating dose (which is usually derived from personal dosimeters) and $H_p(0.07)$ for measuring the shallower penetrating dose [2]. Thus, $H_p(10)$ is sufficient to simulate, for example, the neutron or gamma dose as these penetrate deeper into the tissues, while $H_p(0.07)$ is more suitable for measuring, for example, the beta dose as it does not have the same penetration capability [2]. For this reason, TVO uses $H_p(10)$, as radiation dose measurements with electronic dosimeters are intended for gamma

radiation in particular. It should be mentioned that, in addition to these reasons, the beta dose is also more difficult to measure due to its local penetrating capability, so that, also in the case of TVO, a chest dosimeter may not detect a radiation source located, for example, at the level of the foot.

As a result of radiation in the power plant environments and the subsequent need for the monitoring and control of radiation doses for both the environment and the personnel in that environment, the power plant sites are divided into three areas: 1) controlled area, 2) supervised area, and 3) an unclassified area [14]. The classification of these areas is determined by estimating radiation exposure and the potential risk to personnel [15], [16]. This estimation relies on measuring dose rates and evaluating radionuclide concentrations in the air, as well as surface contamination levels (activity coverage) [6]. In particular, the classification of controlled and supervised areas must consider the nature of the work being carried out in the area and the magnitude of the radiation risk that is inherent to that work [16]. This can be accomplished, for example, with different measurements.

Controlled area is defined as any area where either the dose rate exceeds $3 \mu\text{Sv/h}$, or where spending 40 hours per week in the area could result in a dose greater than 1 mSv per year [6], [14]. In this instance, the work necessitates measures to safeguard against ionising radiation, due to the potential risks of radiation or contamination [16]. *Supervised area* is defined as an area where a personnel's effective dose may exceed 1 mSv per year, or where the equivalent dose to the lens of the eye could exceed 15 mSv per year [16]. Additionally, the equivalent dose to the skin, hands, arms, feet, or ankles may exceed 50 mSv per year in such an area [6], [16]. The area situated beyond the aforementioned zones is classified as *unclassified* and, as a consequence, is not considered significant in terms of radiation protection [6].

Subsequently, the controlled area should be subdivided into at least three distinct zones, employing the same estimations utilized for the initial zoning [6]. In the case of

TVO, the aforementioned zones are designated as green, orange, and red, respectively [5]. The corresponding dose rates for these zones are illustrated in Figure 1.2. This study focuses exclusively on the work carried out in the *controlled area*.

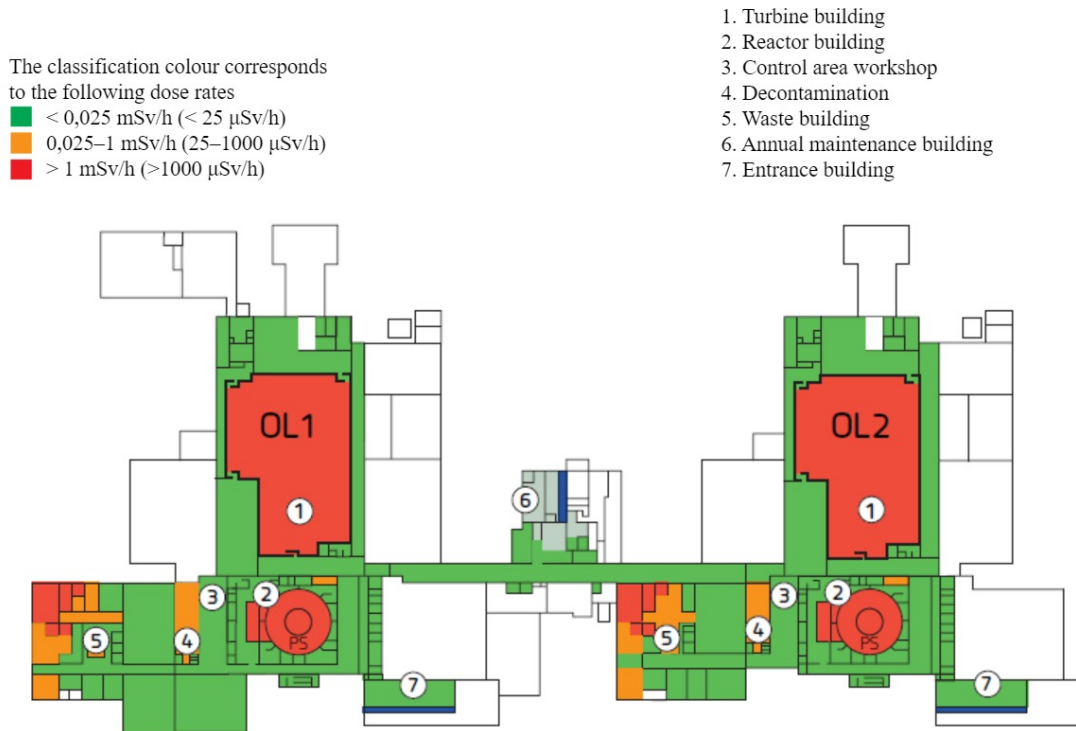


Figure 1.2: OL1 and OL2 controlled area divided into distinct zones [5]

The Figure 1.2 shows the controlled area of the plants during power operation, when the plant is supplying electricity to the national grid. However, this changes as follows when the power plant is in shutdown: areas 1 turn green and areas 2 turn orange. It is essential to state that the data covers work performed during the maintenance periods when these rules apply.

In the event that the work is conducted within a green area, no restrictions are in place and the work may be carried out in accordance with the traditional working hours of 40 hours per week [17]. However, when working in the orange zone, it is needed to plan the work in advance and work areas need to be secured or supervised [17]. Similarly, work in the red zone must be planned and, in addition, carried out

in a short-term manner [17]. It also requires specific dose assessments and detailed planning [17]. Additionally, the work area must be secured or controlled [17].

In the controlled area, the dose and dose rate of individuals are continuously monitored using thermoluminescent dosimeters (TLD) and electronic dosimeters [17], [18]. Electronic dosimeters collect data on a visit-by-visit basis, whereas the data from TL dosimeters are cumulative over a period of time [17], [18]. This means that electronic dosimeters measure the dose for the whole visit and TL dosimeters measure the cumulative dose per month. This latter figure is also reported to STUK as required by law. This study focuses on electronic dosimeters and the data available from them, as this is a more applicable representation of the data in the context of this study.

TVO uses a variety of instrumentation for the quantification of radiation exposure. As was stated, the focus of this study is the electronic dosimeters in use for personal radiation exposure monitoring. In this process, TVO uses Mirion Technology products, the DMC2000S and DMC3000 models [19], [20]. These dosimeters must operate in accordance with STUK regulations, which are based on the provisions required by the Radiation Act (859/2018) [21]. These regulations state that for personal dose monitoring, the unit of measurement is the personal dose equivalent, which we have already discussed, with a maximum measurement uncertainty of 42% [21]. The minimum response range for TLD and electronic dosimeters, specifically for photon radiation (including gamma rays and X-rays), is shown in Table 1.3.

Radiation type	Response range R
Photon radiation (γ and X-rays) $\bar{E}_{\text{ph}} > 10 \text{ keV}$	$0.71 \left[1 - \frac{2 \cdot H_0 / 1.33}{H_0 / 1.33 + H_{\text{ref}}} \right] \leq R$
Photon radiation (γ and X-rays) $\bar{E}_{\text{ph}} \leq 10 \text{ keV}$	$0.5 \left[1 - \frac{2 \cdot H_0 / 1.5}{H_0 / 1.5 + H_{\text{ref}}} \right] \leq R \leq 2$

Table 1.3: Response range for photon radiation with energy greater than 10 keV (kiloelectronvolts), and those with energy less than or equal to 10 keV [21]

In this context, the response R of the dosimeter is expressed as $R = \frac{G}{H_{\text{ref}}}$, where G is the dose determined by the dosimeter, and H_{ref} is the true dose. The parameter H_0 refers to the registration threshold, which represents the minimum dose that the dosimeter records [21]. The symbol \bar{E}_{ph} in the table refers, in this case, to the mean energy of photon radiation, typically denoting gamma or X-ray radiation. Kiloelectronvolt is a measure of energy, where 1 keV equals 1000 electronvolts. This illustrates the dose response and behavior of dosimeters.

In terms of the electronic dosimeters used, the DMC2000S detects gamma and X-ray radiation fields between 60 keV and 6 MeV and operates in the range $1 \mu\text{Sv} - 10 \text{Sv}$; $10 \mu\text{Sv/h} - 10 \text{Sv/h}$, while the DMC3000 is more accurate and detects radiation fields between 15 keV and 7 MeV and operates in the range $1 \mu\text{Sv} - 10 \text{Sv}$; $0.1 \mu\text{Sv/h} - 20 \text{Sv/h}$ [22]–[25]. However, it is also possible to go outside the measuring ranges up to a certain point, as accuracy deteriorates.

Both of these dosimeters measure the personal dose equivalent $H_p(10)$, as mentioned previously and required by the regulations [21]–[23]. In addition, the dosimeters measure the corresponding dose rate [22], [23]. The measurement uncertainty of these dosimeters is also within the limits required by the regulations, ensured by calibration, testing and keeping the internally defined limits below the limits required by the regulations, which in the case of TVO is $\pm 15\%$ [24], [25]. Furthermore, modules for measuring beta and neutron doses are available for the dosimeters, although beta modules are not in use at TVO. Neutron modules are only used when required by the work [23].

These dosimeters also have personal or work dose and dose rate limits that can be used to detect abnormal dose trends at an early stage [26], [27]. These are set via the radiation management system to the lowest visit dose limit from a range of work, daily, monthly and annual dose limits, based on radiation personnel class, gender and work dose code, and if the daily, monthly or annual dose limit is exceeded, the

personnel is prevented from entering the controlled area [26], [27]. Alarm limits are discussed in more detail in Chapter 3.2.1.

Dosimeters are therefore a way of controlling the work and the resulting radiation dose, and also a way of warning the personnel of the dose or dose rate, based on legal requirements and also on economic objectives, since the radiation dose received can also be measured in monetary terms outside the health risk. In this case, the organisation does not need to hire and train additional staff if the radiation dose can be kept as low as reasonably achievable. In the 1990s, it was estimated in the USA that avoiding one millisievert of radiation exposure could save between \$20 and \$2600, depending on the work task [28]. In a 2002 survey in Finland, the value of one man-Sv was estimated to be 77.21 euros, based on the recommended value from 1991 [2].

The aforementioned policies and protection measures are founded upon the premise that work involving occupational exposure should be covered by a system of protection for practices [8]. Consequently, STUK has also defined the previous specifications as requirements in its YVL C.2 regulation, which is in compliance Finnish legislation [6]. The following principles should apply to such work: 1) justification of practices, 2) optimization of protection, and 3) use of dose constraints [1], [8]. The justification ensures that activities involving radiation are only performed if the benefits outweigh the risks [1]. Optimization of protection balances protection levels with economic and social factors, aiming to keep doses *As Low As Reasonably Achievable* (ALARA), ensuring personnel exposure remains minimal during justified work [1], [2], [8]. Finally, dose constraints serve as reference points, typically a fraction of the total dose limit, to prevent excessive exposure. At TVO, this is applied through the ALARA program with set reference levels [1], [27]. Discussion of these methods at TVO is provided in Chapter 3.1.

On the basis of these guidelines, therefore, it must be assumed that radiation doses will accumulate for personnel in any case, because the work and maintenance would be too expensive if no doses were to accumulate for personnel at all. Therefore, planning, design and work should always be based on the ALARA principle of keeping the doses as low as reasonably achievable [3].

The ALARA principle, like the other policies on which TVO bases its operations, is not unique to the power plants operated by TVO, but many, if not all, operators in the nuclear industry base their policies on the same scientifically proven policies, such as the ALARA policy created by the ICRP (formerly ICRP) in the 1977, which is based on the policies created by the ICRP (*International X-Ray and Radium Protection Committee or Commission*) in 1928 [29], [30]. For example, the 1994 US regulations on the application of ALARA guide the ALARA regulations to the corporate policy level, which is supported in practice by the Radiation Protection Manual [28]. In addition, the use of RWP (*Radiation Work Permit*) has long been recognised as a way of controlling procedures during work, such as working time and protective equipment [28], [31]. ALARA is also used to assign responsibilities, whereby the role of the radiation protection manager in an organisation is to enable the practical application of ALARA principles by the engineering organisation [28]. These same policies are also in use at TVO, although their application is at a different level.

In addition to TVO, STUK monitors radiation doses based on TVO's data. Figure 1.3 shows the collective radiation doses received by personnel at Olkiluoto since plant operations began, as reported by STUK [32]. These doses are measured using TL dosimeters and have significantly declined since the early 2000s, thanks to improved technical and radiation protection measures. TVO also monitors environmental radiation and emissions around the plant, though this falls outside the scope of this study [33].

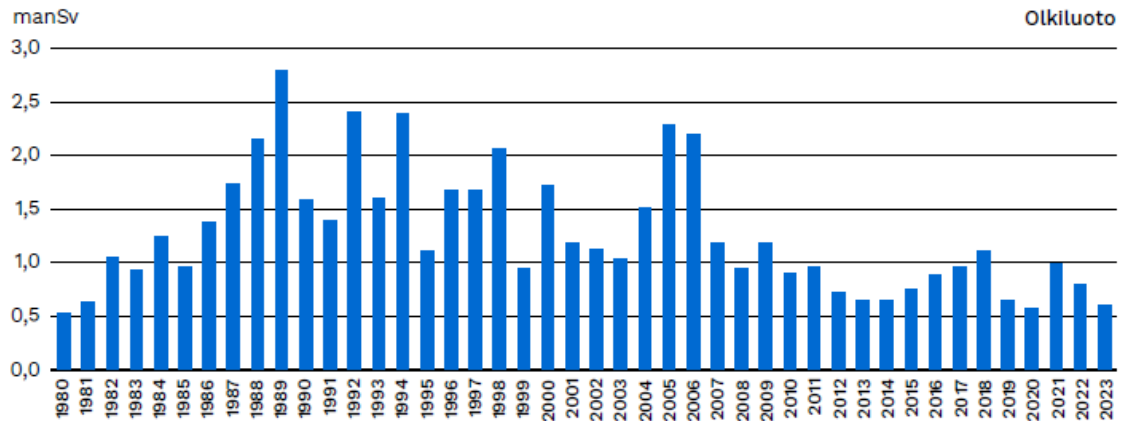


Figure 1.3: Annual collective radiation doses at TVO NPPs [32]

In the following Chapters, this study will focus on external exposure to ionising electromagnetic radiation (γ and X-rays) within the controlled area. This is because the data available for analysis exclusively concern dose data from these types of radiation. However, the biological effects of radiation are treated as a whole, so other types of radiation are discussed here in terms of these effects.

This study builds on the foundations laid in this Chapter, examining the effects of radiation discussed in Chapter 3, and finally reviewing the current radiation-related machine learning applications used in the nuclear power industry. This review forms the basis of the experimental section in Chapter 4, the results of which are discussed in more detail in Chapter 5, which also includes a discussion of the implementation of an appropriate model through to production as part of a network information system solution. The study concludes with Chapter 6, which provides a summary of the research findings. Chapter 3 addresses research question one, while Chapter 4 addresses research question two. Finally, Chapter 5 answers research question three.

2 Research process

The material for this study was obtained during the final two quarters of 2024 through the utilisation of the Volter database, belonging to the University of Turku, in conjunction with the PubMed and Web of Science databases. Furthermore, legislative publications were sourced from the Finlex and Stuklex websites. In addition, this study has also utilised internal sources of TVO, which are not publicly accessible. The aforementioned sources are indicated as unpublished in the bibliography with the specific annotation "unpublished".

In the case of non-refined manual searches, the titles, keywords and abstracts of the materials were targeted, which included books, conference proceedings and articles. The methodology used for the search was comprehensive rather than narrow in order to reach as many relevant publications as possible related to the research topic. The publications were limited to those in English and Finnish and were not restricted by the publisher's Impact Factor (IF) or by publication date, as this would not have had the desired effect due to the historical dependence of the topic.

Most publications were retrieved from the Web of Science database using either an advanced keyword search or a manual match search. A total of 80 publications (N=80) were found for use in this study, most of which were found using search terms related to radiation, its effects, machine learning and regulations. These publications were assessed for their usefulness, but the study did not take a position on whether the publications were peer-reviewed. The high-level search process, inclusion and exclusion criteria are shown in Figure 2.1.

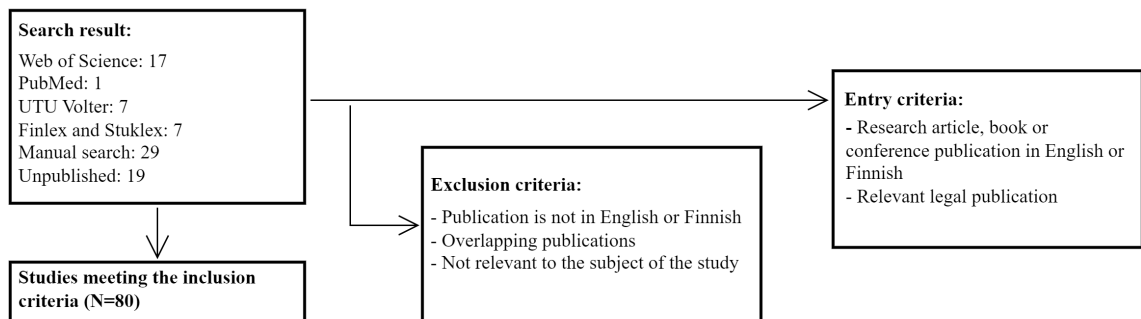


Figure 2.1: Search results that includes a breakdown of sources (manual search results can be from any source)

This research uses existing knowledge, secondary data, to establish the necessary context for experimentation, as no prior implementations of this specific machine learning problem were found in the literature. Consequently, a survey of similar methods applied to machine learning and radiation prediction was conducted, as discussed in Chapter 3.3. The experimental data used in this study is proprietary to TVO and will not be publicly shared.

This study aims to follow the CRISP-DM methodology already mentioned. This process model consists of six steps, the first of which aims to achieve a project understanding, in which it is understood what is being done and with what objective [11]. We have already defined this in Chapter 1. It should be noted that this study is based on a more technical approach than that of a study with an economic or social objective. The steps of CRISP-DM are shown in Figure 2.2.

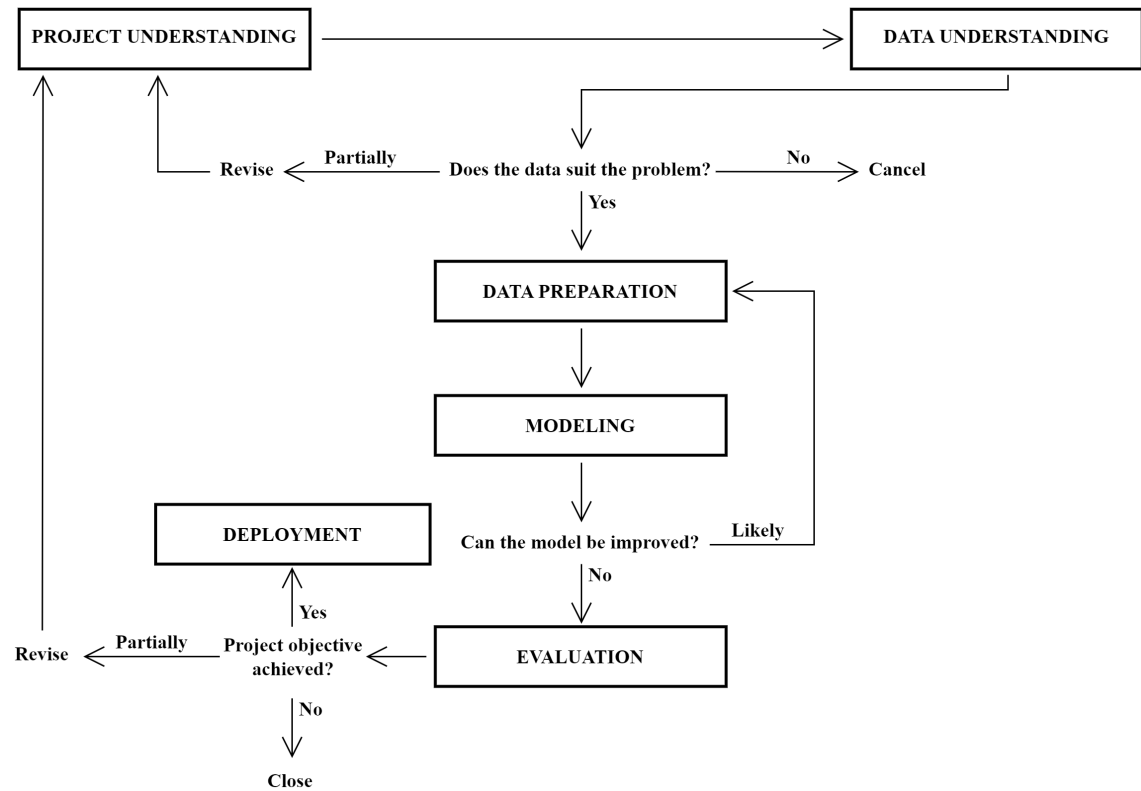


Figure 2.2: CRISP-DM process flow overview [11]

If the flow of this study is compared to the introduced flowchart of CRISP-DM, Chapter 1 corresponds to understanding the project, Chapter 3 to understanding the data, Chapter 4 to data preparation, modelling and evaluation. Chapters 5 and 6 continue with evaluation and introduce deployment as a new aspect.

The CRISP-DM process model was selected because it is the industry-independent de facto standard, despite being published over 20 years ago [34]. Overall, this process model is iterative and flexible, which also means that CRISP-DM follows a natural flow rather than forcing the flow of projects [35]. It is evident that CRISP-DM is not directly intended for machine learning as it fails to consider the fact that machine learning models degrade over time [35], [36]. Consequently, post-deployment operations must also be taken into account, as discussed in more detail in Chapter 5.2 [36].

3 Radiation and machine learning

This research adopts an exploratory approach to understand the data and the issues associated with it. The meaning and origins of the data are discussed from the perspective of what policies and regulations they are based on, i.e. why something is done as it is done. In addition to creating understanding, we establish the foundation for these policies and regulations, one of which is to minimise the health risks associated with radiation. The analysis of these health risks represents a significant aspect of this study, which will also facilitate the reader's comprehension of the sensitivity of the data and the subject matter.

The following Chapters will address the data, subsequently examining the nuclear power plant environment and the processes through which data is formed. We then proceed to analyse the effects of radiation doses, before concluding with a discussion of the current applications of machine learning in a similar context, along with the general understanding of machine learning and the models utilised in this study.

3.1 Legislative approach to dose regulation

As previously noted, TVO's operations are subject to numerous radiation-related regulations and directives, which are overseen by STUK [6]. Similarly, in radiation protection and radiation-related work, the principles are set forth in the Radiation Act (859/1987), the Nuclear Energy Decree (161/1988), the Nuclear Energy Act (990/1987), the Government Decree on Ionising Radiation (1034/2018), and the

Ministry of Social Affairs and Health Decree on Ionising Radiation (1044/2018) [15], [16], [37]–[39]. In terms of this study and in accordance with YVL C.2 regulation, STUK oversees the practical safety measures on radiation protection of personnel and monitoring of radiation exposure in nuclear facilities [6]. It is important to note that while STUK now enforces radiation safety protocols, the original design and operation of nuclear power plants in Finland were based on US requirements from the 1970s [40]. Over time, these requirements evolved through national and international cooperation via laws and regulations [40]. For instance, early safety requirements did not mandate preparedness for severe accidents [40].

Under the Radiation Act (859/1987), radiation protection is based on the principles of justification, optimisation and protection of the individual [6]. These can be seen as an interpretation of the way in which we have already discussed the protection of work that may involve ionising radiation, using a system of protection for practices [27]. TVO applies these policies by defining an ALARA programme, of which this study focuses on OL1 and OL2 and dose limits [27]. For occupational doses, TVO has set an internal limit of 10 mSv per year, which acts as a maximum annual dose instead of the 20 mSv limit stipulated by law, for effective doses received by personnel [27].

TVO has identified in its ALARA programme a number of ways to reduce doses, such as preventing and controlling fuel leaks, source term minimisation, decontamination, work planning and development of work methods, daily dose limits and work codes in the work dosimetry system, a risk-based rate control programme, contamination control and annual maintenance [27]. Of these, this study addresses the data-driven quantifiable attributes which are the daily dose limits and work codes of the occupational dosimetry system. Other methods are directly reflected in the resulting doses through the implemented practices.

The legislation and the resulting STUK regulations limit the intake of radiation doses so that the annual effective dose does not exceed 20 millisievert (mSv) per personnel [6], [7], [38]. In practice, it is not possible to get anywhere near such a dose, in part because of the lower dose limits in force. The aforementioned dose limit is also affected by the radiation classification of a personnel exposed to radiation, i.e. class A and B [6], [7], [41]. For a personnel classified as class A, the effective dose may exceed six millisieverts per year (15 mSv for the eye and 150 mSv for the hands and feet) [6]. If this is not the case, the personnel falls into class B [41]. Typically, class A workers are employed in roles such as radiation protection technicians or reactor operators [41]. For example, the threshold for considering a change in personnel classification from B to A class is 3.5 mSv among other requirements [6]. In addition to these dose limits, TVO has defined its own period and work-specific limits, as previously referenced. These limits are overseen by the radiation management system, with the work dose code (WDC) serving as a central component of this management mechanism. These limits will be discussed in the Chapter 3.2.1 and represent a significant component of the data set analysed.

3.2 Nuclear power plant setting

From the point of view of this study, data collection starts as soon as the required work have been defined for a given annual maintenance. In this case, at the specified time, the personnel enters the controlled area of either the OL1 or OL2 plants (site visit) to carry out the work. Before doing so, the personnel logs in with their electronic dosimeter using the work dose code assigned to the work. The personnel will then carry out the work on the defined system. In this case, as soon as the work itself, whether completed in one visit or not, have been completed, the personnel leaves the controlled area and signs their electronic dosimeter out. This constitutes a

single site visit, which always has one dose as measured by the electronic dosimeter. A single site visit may therefore include several different work tasks under the WDC.

The data used in this study span over annual maintenances from April 4, 2012 to May 22, 2023. The starting date of 4 May 2012 was chosen because, in the two years prior, the outputs of the OL1 and OL2 plants were each increased by 20 MW. Additionally, the plants were upgraded with new low-pressure turbines, generator cooling systems, seawater pumps, and internal isolation valves for the main steam pipes [5]. The final data sample was obtained on 31 May, 2023 as the year 2024 was left out for testing purposes. Overall, between 2012 and 2023, there have been hundreds of thousands of site visits during the annual maintenance periods, with each visit representing a single data point from a raw data perspective. The use of automated and access-management-dependent data collection processes has resulted in the near complete data set, with the exception of a few observations, which are addressed in the next Chapter 3.2.1.

3.2.1 Site visit formation

As previously stated, the dose received by personnel is subject to constraints depending on different time intervals. To further illustrate this point, a male A-class personnel may receive a dose of 1.5 mSv per day during the annual maintenance period, for example, assuming the annual internal limit of 10 mSv is not exceeded [26]. It should be noted that this calculation does not take into account any additional constraints, such as those specified by the WDC, which limits the dose by a per-visit dose limit. WDC is required to enter the controlled area of the plants and is a key element of data and radiation dose monitoring. Personnel uses this code to log in to the controlled area to carry out their work.

WDC is made up of three parts: 1) plant ID, 2) system ID and 3) project ID. WDC is therefore given in the form XYYYYZZ, referring to the previous structure [42].

For example, in the case of the Olkiluoto 1 plant, system 100 and project 01, the code would be 110001. Using this system, the dose limit can be set on a work basis, thereby allowing for better dose control [42]. This is further complemented by a dose rate limits, which can also be utilised to alert the personnel if the predetermined limit is exceeded. If no maximum dose or dose rate limit is set for the WDC, the default values are used [26], [42].

In this thesis, the associated dose and alarm limits will be treated as features, with WDCs being parsed in order to identify the specific plant and system being worked on, as well as the corresponding project. In addition to visit-related data, this study utilises information regarding the personnel involved. This includes the personnel's radiation class, the completion status of mandatory radiation work training (e.g. entry-level and advanced training when required), and whether the personnel is employed as a subcontractor or not. Each site visit to a controlled area is linked to a specific time, thus enabling the dataset to be observed as a time series. The time spent on each visit can be derived from these time-stamped records. We will be discussing more about the data and how it can be processed for the machine learning task in Chapter 4.1.

3.2.2 Radiation phenomena and dosage

As briefly touched upon previously, electromagnetic radiation is a collective term for a group of radiation types, which includes γ (gamma) and X-ray (X-ray) radiation. Unlike other forms of radiation, these are measured on a per-visit basis when entering the controlled area at NPPs using electronic dosimeters and are therefore examined in this study. This Chapter will provide a more detailed discussion of these types of radiation, which will then be used as a point of reference in the next Chapter when exploring the biological effects of radiation and the reasons behind its detrimental effects on the human body.

Ionising radiation can be classified into two main categories: charged particles and *neutral radiations* [3]. The latter group includes gamma rays, X-rays and neutrons, of which neutrons are not considered in this study because this is not measured in terms of the data used. Unlike charged particles, neutral radiations do not directly ionise matter through Coulombic interactions, meaning the repulsive forces between two charged particles based on the distance and electric charges between them, as neutral radiations lack the needed electric charge (energies in question are limited to a few MeV) [3]. Instead, they interact with matter primarily through indirect processes, including: 1) Compton effect, 2) photoelectric effect and 3) pair production [2], [3]. These result in indirect ionising radiation [2].

The Compton effect, also referred to as photon-electron scattering, represents a process whereby a photon of energy equal to its rest mass energy, $E = h\nu$, interacts with an electron of rest mass m_e , where E represents the energy of the photon, h Planck's constant and ν the frequency of the photon [2], [3]. Despite the electrons in an atom being bound to the nucleus, their binding energy is considerably less than that of typical gamma rays. As illustrated in Figure 3.1, the photon is deflected and loses energy in the collision, becoming a photon of new energy $E' = h\nu'$ and the electron gains energy and moves away from the atom leaving it ionised [2], [3]. This collision depends on the photon scattering angle θ and the principles of conservation of energy and momentum [3]. In this process, the greatest energy loss by the photon occurs when it is scattered backwards (at an angle of 180°) [3]. In the event that E is considerably greater than the rest energy of the electron, the final photon energy, designated as E' , is approximately half of the initial energy [3]. Compton effect is the most dominant at energy range of 0.1 MeV to 10 MeV [2].

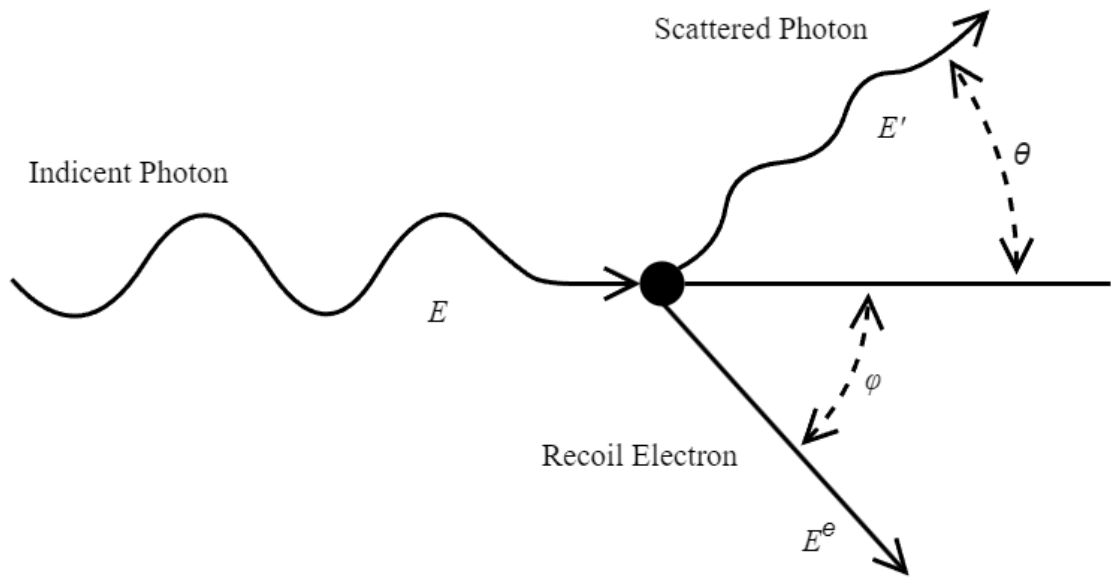


Figure 3.1: The Compton effect: photon scattering results in energy transfer to an electron, with the photon losing energy and being deflected. The atom is left ionised [2], [3]

Photoelectric effect (photon–electron) is an interaction, whereby a photon transfers all of its energy to an electron, which is then ejected from the atom and loses its energy, leaving a positively charged ion and ionising the medium [2], [3]. In this instance, the photon is absorbed, resulting in the ejection of an electron, which is known as a photoelectron (Figure 3.2) [2], [3]. This photoelectron gets ejected with an energy of $E_K^e = E_\gamma - I$, where E_γ is photon's energy, I is the potential for ionisation of the electron to the atom [3]. The photoelectric effect is dominant at lower photon energies and is only possible in the energy ranges of tens to hundreds of kiloelectronvolts (keV) [2]. Both the photoelectric and Compton effect processes may provide electrons with sufficient energy to ionise other atoms [3]. Additionally, following the ejection of an electron, light emission or X-ray production may occur [3].

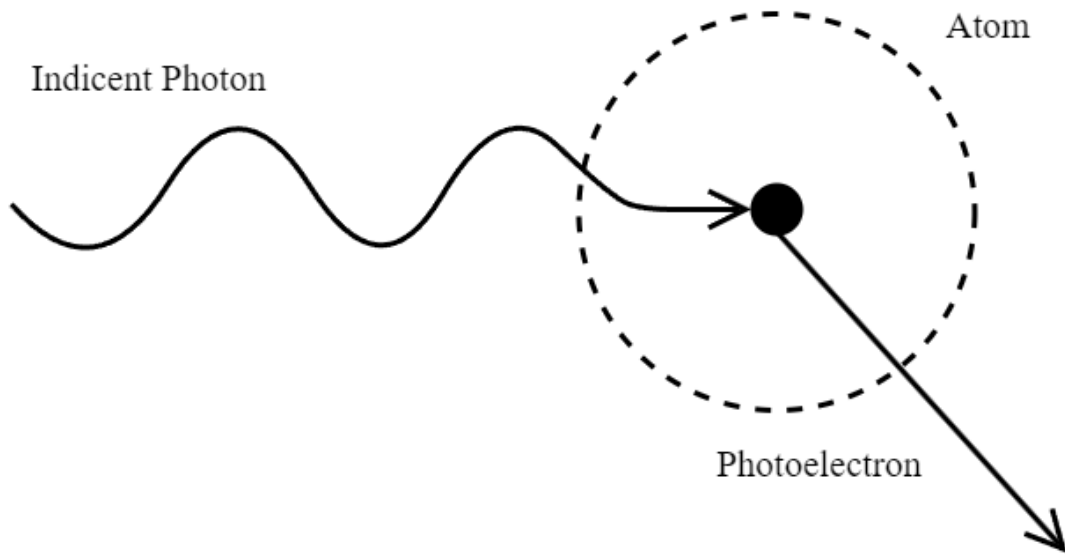


Figure 3.2: Photoelectric effect: energy is transferred from a photon to an electron. This results in the ejection of the electron from the atom as a photoelectron, and the creation of a positively charged ion [2], [3]

In pair production and annihilation the photon energy is converted into mass, typically in the vicinity of a nucleus due to the presence of a strong electromagnetic field [3]. This only occurs when the photon energy exceeds 1.022 MeV [2]. In pair production, a photon interacts with the Coulombic field of a nucleus, resulting in the creation of an electron (negatron) and a positron (Figure 3.3) [3]. The energy of the photon is transformed into mass, with the total mass equalling the combined mass-energy of the electron and positron, which is 1.022 MeV [3]. Any photon energy in excess of 1.022 MeV is distributed as kinetic energy to the electron and positron, in accordance with the equation $E_K^{\text{pair}} = E_K^{e^+} + E_K^{e^-} = E_\gamma - 2E_0^e$, where E_γ is the photon energy, $E_K^{e^+}$ and $E_K^{e^-}$ are the kinetic energies of the positron and electron, and E_0^e is the rest mass energy of the electron [3].

Once the positron has lost its kinetic energy, it rapidly combines with a nearby electron in a process known as annihilation [3]. This process results in the complete annihilation of both particles, accompanied by the emission of two gamma photons,

each with an energy of 0.511 MeV [3]. The emission of these photons occurs in opposite directions, resulting in the conservation of momentum through the transfer of photon energy [3].

In both pair production and annihilation, high-energy particles, such as electrons and positrons, possess sufficient energy to ionise atoms by ejecting electrons from their outer shells as they traverse matter [2]. This process results in the production of ionising radiation, which can subsequently cause further ionisation events as the particles lose energy through collisions [2]. The gamma rays emitted during annihilation also contribute to the overall level of ionising radiation, as they can interact with matter through indirect processes, as previously outlined, which further displace electrons from atoms and lead to ionisation [2].

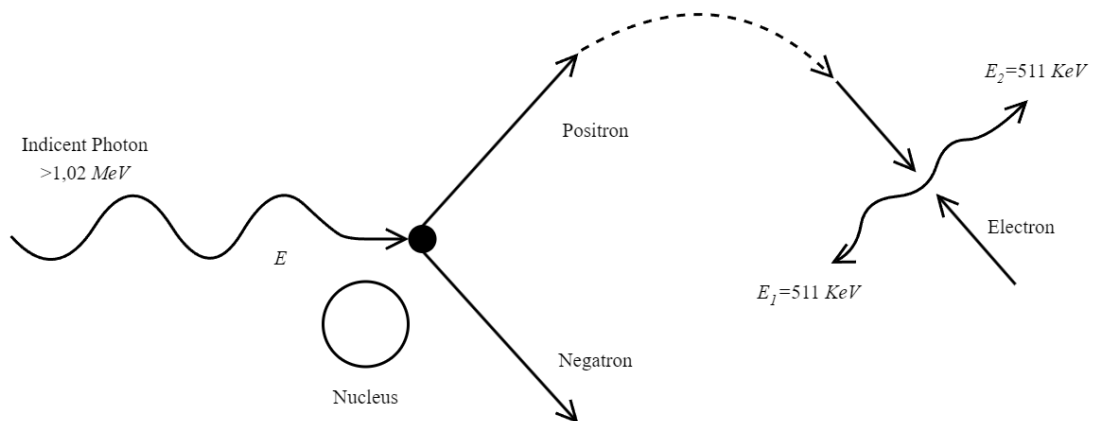


Figure 3.3: Pair production and annihilation: an photon interacts with a nucleus, producing an electron and positron. The positron eventually annihilates with an electron, releasing two 511 keV gamma photons in opposite directions to conserve momentum [3]

Due to these indirect processes, as well as the ability of other types of ionizing radiation to disturb molecular bonds, the following bonds are most affected, in order: 1) metallic bond (least), 2) ionic bond and 3) covalent bond (most) [3]. Of these bonds, covalent bonds are the most abundant in biological tissues and are therefore the most vulnerable to ionizing radiation [3]. This can lead, among other things, to direct damage to cellular molecules, such as DNA, and indirectly through

chemical reactions that can further damage biological structures [3]. Electronic dosimeters therefore measure these effects and model the quantifiable radiation dose to the human body. In the data, the resulting radiation dose, the predicted value in classified form, is expressed in microsieverts μSv .

3.2.3 Biological effects of radiation

Radiation can have a wide range of biological effects, which are typically classified into two main categories: 1) deterministic and 2) stochastic effects [8], [43]. In addition to these, physiological effects can be classified as 1) somatic, 2) genetic and 3) teratogenic [3]. These are combined in that deterministic effects are rapidly apparent in cell death or malfunction in clinical manifestations, while stochastic effects are irregular or statistical in nature, with somatic differentiation referring to the body or its condition, genetic referring to genes that enable gene inheritance, and teratogenic referring to fetal and embryonic development [2], [3], [8], [43]. From the perspective of this study, we are examining the *genetic effects from a stochastic standpoint*, as these are the most relevant in the context of a nuclear power plant environment. These stochastic effects are expressed in the ways shown in Figure 3.4, either through carcinogenic cell division or erroneous repair, possibly affecting offspring through chromosomal mutations [8]. While the effects of radiation can also be deterministic, this is not within the scope of our focus, as it is not a probable occurrence, despite the theoretical possibility. However, as a reference value, a sudden dose of 0.2 Sv (200 mSv) does not show a somatic clinical effect, but a dose over 4 Sv, for example, is likely to result in death without treatment [3]. For example, at the time of the Chernobyl nuclear accident, approximately 134 workers and firefighters received doses of 0.7 Sv (700 mSv) to 13.4 Sv (13400 mSv), resulting in a reported 28 deaths from direct exposure [44].

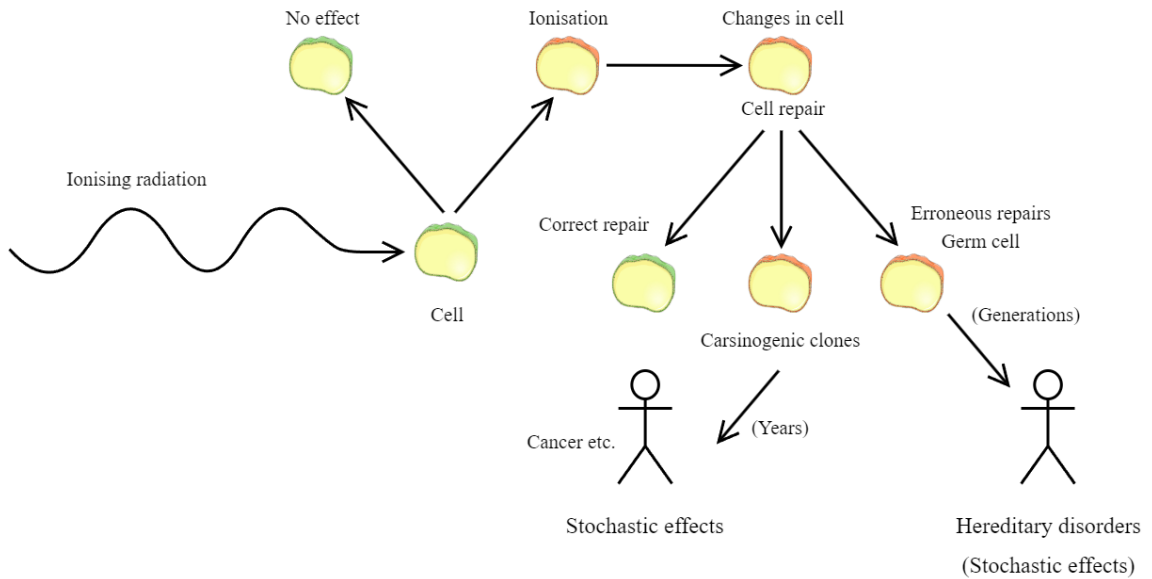


Figure 3.4: Overview of the stochastic effects of ionising radiation (not including deterministic effects) [2], [43]. Icons emptycell-3d-2 and emptycell-3d-3 by Servier are licensed under CC-BY 3.0 Unported

The literature has identified a correlation between radiation dose and an increased risk of chromosomal and chromatid abnormalities [45]. These abnormalities act as early markers of stochastic genetic effects, such as cancer [45]. It has been evidenced that DNA damage responses in individual cells are significant for the development of cancer cells, in addition to gene and chromosomal mutations, even at low doses [43]. In particular, it has been identified that high dose and dose rate positively correlate also with non-cancer mortality [46]. Chromosome analyses have demonstrated that chromosomal defects, particularly in individuals exposed to long-term low-dose radiation (LDR), are associated with increased genomic instability [45], [47]. Studies have also shown that occupational exposure to LDR (a few hundred mSv) can result in chromosomal aberrations especially in shorter-term [45]. The LDR in question represents, in some cases, the typical dose received by personnel in their lifetime in the context of nuclear power plants (NPPs) [4], [45]. Nonetheless, it has been found that even when the effective doses do not exceed the permissible limit of 20 mSv per year, personnel in nuclear power plants exposed to ionising radiation exhibit

significantly higher levels of chromatid and chromosome aberrations compared to control groups [45].

Additionally, a statistically significant correlation has been identified between radiation dose and non-cancer mortality, particularly in relation to cardiovascular disease [46]. There has also been reports of radiation exposure and a link to general circulatory diseases [46]. This indicates that cancer is not the sole risk factor, although it is frequently associated with the stochastic effects of radiation. Furthermore, elevated levels of stress, anxiety, and post-traumatic stress disorder (PTSD) have been observed among personnel and civilians following nuclear accidents [48]. Such psychological stress may indirectly increase susceptibility to stochastic effects.

Furthermore, it has been found that in radiation-exposed family triads where father was occupationally exposed to radiation (mean lifetime gonadal dose from gamma radiation 1.65 ± 0.08 Sv) experienced heightened mutation frequency, evidenced by an increased frequency of chromosomal abnormalities [47]. It is important to acknowledge that these effects cannot be fully assessed in isolation, as lifestyle factors also contribute to the overall picture. However, the existing literature tried to account for these variables.

As previously stated, LDR remains a concern for plant personnel with regard to genetic and epigenetic alterations, even at relatively low doses. In South Korea between 2012 and 2021, despite average individual doses remaining well below regulatory limits, approximately 0.39 mSv per year, a fraction of personnel received higher doses in specific work tasks [49]. It is additionally noteworthy that typically majority of personnel receive a minimal radiation dose, if any, but a small number receive significantly higher doses in certain occupations [49], [50]. This highlights the importance of the issue in certain work settings and the data imbalance that should be considered when processing the data for this study [49]. However, the results of

all studies are not directly comparable, because for example, Korea had an annual effective dose limit of 50 mSv in 2023 compared to Finnish 20 mSv per year [6], [49].

At this time, taking into account the literature, it must be stated that there is an insufficient amount of data available to conduct a comprehensive statistical analysis of the effects of radiation, although we can identify chromosomal abnormalities [2], [3], [47]. This is due to the ethical concerns associated with deliberately exposing humans to harmful radiation levels for experimental purposes. The majority of available data are derived from incidents and medical or occupational exposures, which are insufficient for comprehensive statistical analysis [3]. A substantial proportion of the historical data is drawn from the aftermath of the nuclear bombings in Japan in 1945 [2], [3].

However, as stated in the 2007 ICRP publication, even with exceptions, the cellular processes and dose-response data support the premise that within low-dose ranges (below approximately 100 mSv), the likelihood of cancer or heritable effects increases in direct proportion to the equivalent dose in relevant organs and tissues, even though there is an absence of direct evidence that radiation exposure leads to heritable diseases in offspring [43]. Nevertheless, the ICRP has reached the conclusion that there exists compelling evidence to suggest that ionising radiation can result in heritable effects in experimental animals [43].

In the context of radiation effect models, the Linear No-Threshold (LNT) model is regarded as the most conservative approach. LNT model assumes that the risk of cancer or other harmful effects increases in a linear fashion with dose, even at very low exposure levels, extrapolating to zero exposure [2], [3]. This model is also in use at TVO. However, this model has been the subject of considerable criticism on the grounds that it lacks direct empirical evidence, particularly in relation to low-dose exposures [3]. Alternative models, such as hormesis suggest that there may be a certain level of exposure below which radiation might not cause harm [3]. However,

these models are also controversial and lack scientific consensus. Therefore, it is generally assumed that no specific threshold exists below which radiation can be considered entirely harmless, while there is some estimates that doses under 500 mSv does not cause genomic instability over generations [2], [3].

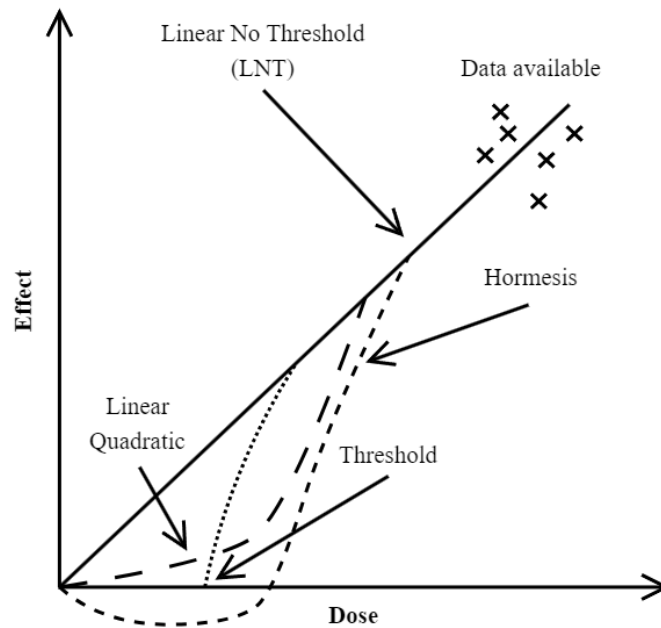


Figure 3.5: Radiation effect to dose models. Available data does not conclusively support effects in the low-dose range [3]

3.3 Related studies and machine learning

This Chapter discusses the various machine learning implementations in nuclear power plant environments that can be found in the public literature because, as mentioned, no similar implementation of radiation monitoring using machine learning could be found in the perspective of this study. The purpose of this Chapter is to establish an understanding of what implementations exist, what models have been used and what methods have been used to evaluate them. This provides the context for the following Subchapter, 3.3.2, which outlines the definition of machine learning, the steps involved and the machine learning techniques used in this study.

First, however, it is important to note that machine learning differs significantly from the current, more traditional way of predicting received radiation doses. Current manual prediction requires expert knowledge and manual work, which is also based on trends and dose results from previous work, which can be difficult to process due to the large amount of data. Machine learning, on the other hand, is more efficient at handling large amounts of data and can detect non-linearity between variables, meaning it can uncover complex patterns and relationships that traditional methods may miss, leading to more realistic predictions. This is also reflected from the fact that the MAPE value already mentioned was 18.76% since the last power increase in 2012. A sufficiently accurate model can therefore reduce the human tendency to make mistakes in a data-driven way.

3.3.1 Present-day applications

This study will review seven existing studies that have used machine learning to predict radiation levels, although the implementations of these studies are not directly comparable to the approach taken in this study. These seven studies focus primarily on nuclear power, as these were selected based on the relevance to this research. However, there are also examples of radiation prediction using machine learning in the hospital and pharmaceutical industries.

The studies in question address either the reactor itself or the reactor building, in addition to environmental radiation monitoring utilising machine learning. The most closely aligned approach to the subject of this study was the estimation of the personal dose equivalent using photon energies measured by TL dosimeters, which was also successfully accomplished [51]. Other applications of machine learning have been utilised, for instance, in the aftermath of the Fukushima disaster to predict the ambient dose rate and in Germany for anomaly detection using the gamma dose rate, taking into account environmental conditions [52], [53]. Additionally, in

the field of reactor engineering, stress intensity in reactor pressure vessels and the impact of radioisotopes in primary coolant loops have been modelled to examine the influence of corrosion products, with the objective of predicting and minimising radioactive corrosion levels and, consequently, reducing radiation dose levels at the reactor shutdown [54]–[56]. An alternative approach was the implementation of a monitoring system utilising machine learning to monitor radioactive materials in nuclear facilities. This system employed a sensor network to model the tracking of radioactive sources and identify nuclide types. This system could be used, for instance, in storage facilities for the protection of waste in the event of theft [57].

It is worthy of note that existing machine learning solutions do not inherently utilise data that is linked to personnel working in NPPs. Therefore, the methodology proposed in this study, which entails modelling radiation doses in a site visit oriented manner based on the nature of the visit, represents a distinctive approach to predicting the radiation doses associated with occupational activities. In exploring suitable machine learning models for this methodology based on the reviewed studies, we observed that Decision Trees (D-Tree), Random Forests (RF) and Artificial Neural Networks (ANN) were the most common models used [51]–[53], [57]. In addition to these, modelling was also done with Gaussian Bayesian Networks (GBN), Gaussian Processes (GP) and Light Gradient Boosting Machines (LightGBM) [54]–[56]. A detailed explanation of the models used in this study can be found in Chapter 3.3.2, and therefore, they will not be discussed in this Chapter.

Approximately half of the studies discussed applied some form of cross-validation for evaluation purposes [54], [56], [57]. In particular, regression-type problems were evaluated using Root Mean Squared Error (RMSE) or Negative Log-Likelihood (NLL) in several cases [52], [55], [56]. In other instances, accuracy and a type of confusion matrix implementation were common [51], [53], [54], [57]. Additionally,

hyperparameter optimisation was utilised for the purpose of optimising the model's performance, for example, through the use of grid search [54].

In conclusion, the available literature demonstrates a number of different applications of machine learning in the context of nuclear power, specifically in the prediction of radiation levels and doses. However, there is a notable absence of methods that incorporate data directly linked to personnel activities in NPPs. This study addresses this gap by introducing a distinctive site visit based modelling strategy to predict radiation doses associated with site visits.

3.3.2 Overview of machine learning

In addition to the supervised machine learning models already mentioned, it should be noted that supervision is one of the main approaches of machine learning, along with unsupervised learning and reinforcement learning [9]. This study specifically focuses on supervised machine learning (hereafter referred to as machine learning). This differs from other machine learning approaches in that the training data contains the correct answers (targets) to some input data, in the case of (x_i, t_j) , where x_i are the inputs and t_j are the targets indexed by i, j [9]. Input i runs from 1 to the number of input dimensions m ja target j runs from 1 to the number of output dimensions n [9]. The point of the output y (y_j , where j runs from 1 to the number of output dimensions) in this case is to produce predictions based on the input data, which is then compared to the target values to assess the model's accuracy [9]. This leads to the main purpose of machine learning, which is to generalise, i.e. to create sensible outputs from inputs that have not been observed before. In the case of classification, we consider input vectors, from which we decide to which, in this case discrete class N , the vector belongs [9].

Comparing the data preparation, modelling and evaluation phases of CRISP-DM presented in Chapter 2, these phases can be described in more detail from the

perspective of the machine learning classification system shown in Figure 3.6. It is important to note that these stages are not independent, but rather interrelated, and that some methods combine these stages, for example, feature selection and classifier design can occur together [58]. In the following sections, we will provide a detailed discussion of these steps and explain what typically happens in practice, helping the reader gain a clearer understanding of the overall process.

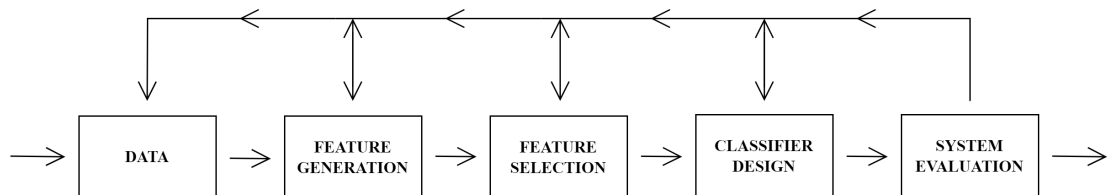


Figure 3.6: Design stages of a machine learning classification system [58]

In this study, the term data refers to our time-series bound features (input and targets) discussed in Chapters 3.2.1 and 4.1. Once the data has been collected, we want to further refine it by modifying it into new features that are more compact and informative than the original features, often through mathematical transformations, domain-specific knowledge or combining features [58]. This is called feature generation. This will usually lead to a reduction in dimensionality and increase the overall quality of the data, even though no feature selection has been made [58]. Overall, feature generation makes the machine learning algorithms identify relevant patterns more effectively and reduces the computational complexity [58].

The objective of the next phase, feature selection, is to select the most important of these generated features in order to reduce the dimensionality of the input, while ensuring that the input retains as much of its class discriminatory information as possible [58]. The aim is to obtain features that exhibit a large between-class distance and a small within-class variance [58]. This implies that the between-class values should be as distant as possible, while the within-class values should be as

close to each other as possible. This objective, along with the most informative features, is achieved through the data preparation techniques and model feature importances outlined in Chapters 4.1 and 4.3 respectively. Explicit feature selection is not a prerequisite for the tree-based models used in this study and thus it is not implemented, given these models are inherently capable of selecting the most informative features. Nevertheless, most informative features will be assessed during the evaluation process.

The next step is choosing and designing the classification algorithm (*classifier*), which depends on the available data [11], [58]. In this study, we use a multi-class classifiers because the target features have more than two possible outcomes. Optimizing the model's hyperparameters, which are specific to each model, is also important [11]. For example, a hyperparameter could be the maximum depth of a decision tree. These hyperparameters affect how the model learns from the data and how it behaves. In the Chapter 4.2, we will apply these machine learning techniques to the collected and preprocessed data.

Finally, one of the most important steps is model evaluation, which determines whether the model has actually learned from the data. This objective can be achieved through the implementation of three different methods: 1) the resubstitution method, 2) the holdout method, and 3) the leave-one-out method [58].

In this study, holdout method is used. This approach was chosen because the amount of data applicable in this study would be too computationally expensive to cross-validate. In this study, we used the holdout method to divide the data into three parts: 1) the training data, 2) the validation data and 3) the test data in a 60/20/20% split [59]. In contrast, the leave-one-out method would have allowed us to use the entire dataset for both training and testing, but this would not have been feasible in terms of the time required to train and test the models [58]. It is evident that the holdout method's limitations are twofold. Firstly, it restricts the amount

of training and test data that can be utilised, as these cannot be mixed, which can affect the models capacity to generalise [59]. Secondly, it is challenging to determine the optimal split between the training and test sets. Allocating a higher proportion of the data to the training set may enhance model accuracy by reducing excess mean error and variance associated with finite datasets [58], [59]. However, this can result in a reduced test set, which can compromise the reliability of evaluation metrics due to increased variability in error estimates [58]. Alternatively, a larger test set may provide more accurate performance estimates, but at the expense of reduced training data, which can potentially increase the classifier’s overall error probability [58], [59].

In the following discussion, the models utilised in this study will be outlined, along with the reasoning behind their selection. These models are presented in Table 3.1. A subsequent general overview will be provided of the theoretical basis of each model, together with an outline of their principles, features, and their relevance in addressing the challenges of this study. These include the already mentioned data skewness, resulting in a low number of data points in high dose rates, and the fact that the problem definition is a multiclass task.

Model	Type	Algorithmic approach
Random Forest	Classifier	Ensemble of Decision Trees
Balanced Random Forest	Classifier	Ensemble of Decision Trees
LightGBM	Classifier	Gradient Boosted Trees
XGBoost	Classifier	Gradient Boosted Trees
Easy Ensemble	Classifier	AdaBoost Ensemble

Table 3.1: Overview of models used in this study, classification types and algorithmic approaches [60]–[64]

Random Forest is an ensemble learning method that builds more traditional Decision Trees and combines their outcomes through majority voting into a single result [62]. These trees are built by randomly sampling the data through bootstrapping

and introducing only a subset of features at any given sample, allowing the remaining data to be used for Out-of-bag (OOB) sample model estimation [62]. The model also allows the importance of features to be examined by examining how much a given feature reduces impurity across all trees by measuring *Gini impurity*, which quantifies how much classes are mixed within a given tree node [62]. Random Forest model was selected as the primary model for this study due to its ease of use and the fact that the *Law of Large Numbers* (LLN) dictates that the model consistently converges and does not lead to overfitting issues [62]. Consequently, the predictions of the Random Forest model are known to stabilise as the number of trees increases, thereby ensuring consistent results [62]. This characteristic, when combined with capacity to manage large datasets makes Random Forest a suitable option for our initial modelling. Random Forest's high-level logic is presented in Figure 3.7.

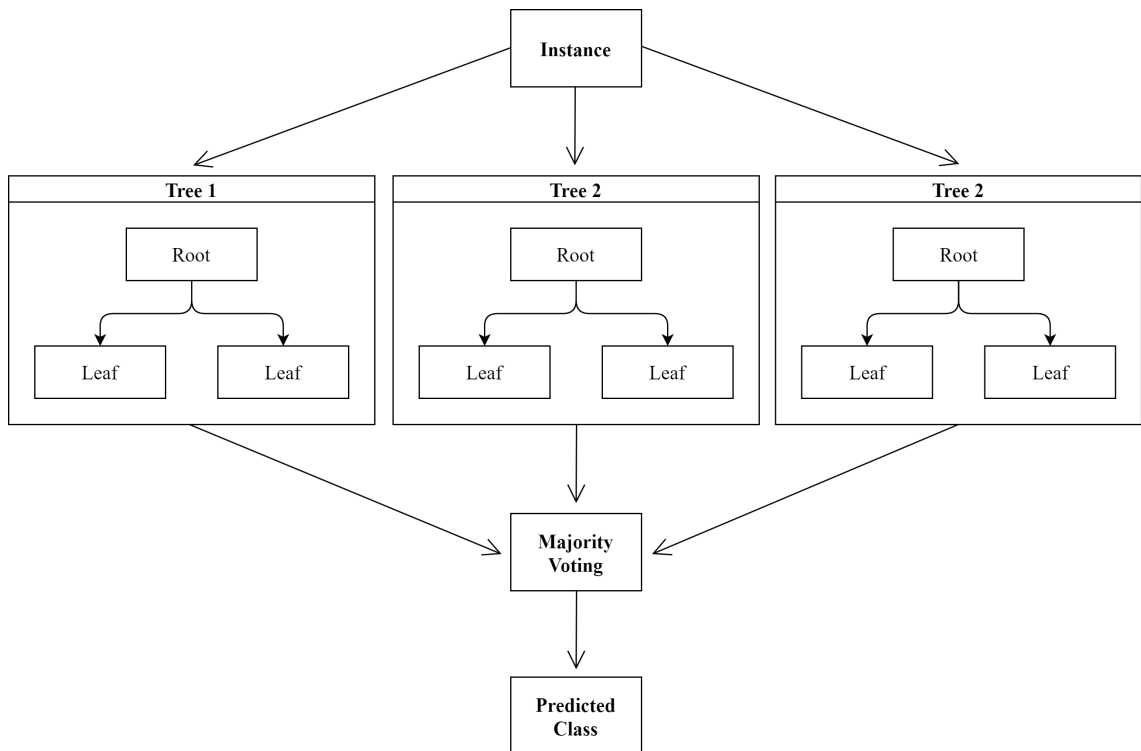


Figure 3.7: Random Forest simplified. Majority vote of decision trees decides the class for the data instance [62]

The Random Forest model was further explored through the application of the Balanced Random Forest (BRF) model, which, while aligning with the principles of the Random Forest, features an alternative approach through the implementation of bootstrap sampling for the minority class with replacement sampling the same number of samples from a majority class as bootstrapped from the minority class [63]. In vanilla Random Forest bootstrapping there is no special consideration for class balance [62]. This approach was taken to address our data imbalance.

Extreme Gradient Boosting (XGBoost) is used in this study because, as an ensemble learning method, which builds a strong classifier from an ensemble of weaker classifiers aligns well with the imbalance and size of our dataset [60]. By sequentially building decision trees where each tree corrects the errors made by the previous ones and with the use of a gradient descent approach to minimize the loss function, leveraging both the first and second derivatives of the loss for efficient optimization, XGBoost suits well to our data imbalance challenges [60]. During the model's training, XGBoost computes a weighted score for each tree split based on a gain metric, selecting the best splits for use and uses a level-wise tree growth strategy, where trees are expanded level by level to maintain balance and improve efficiency [60]. Regularisation is used to smooth out the latest learned weights, shrinkage to scale new weights and subsampling, which is also used in the more traditional Random Forest to prevent overfitting [60]. This makes XGBoost very sophisticated method for a such machine learning problem, but at the same time requires more computing power. XGBoost's level-wise growth strategy is presented in Figure 3.8.

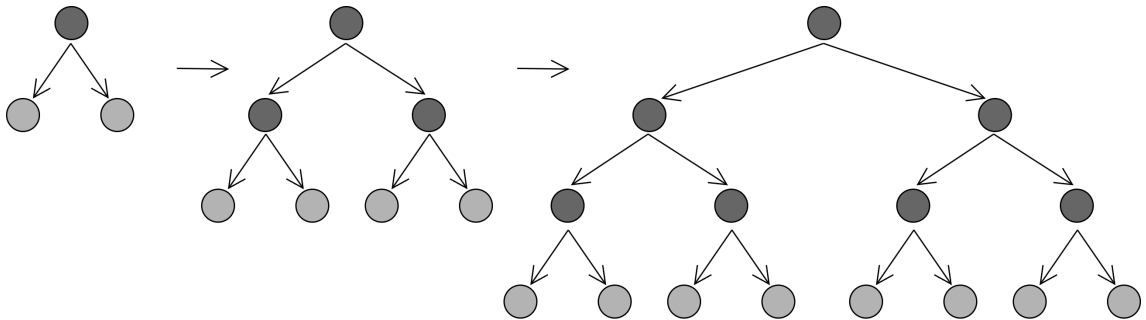


Figure 3.8: XGBoost’s level-wise growth strategy visualised. Black indicates terminal nodes, leaves, while grey indicates the node selected to grow next

LightGBM is an advanced version of gradient boosting, building on the strengths of ensemble methods such as Random Forest mentioned before [61]. The LightGBM’s leaf-wise growth strategy differs from the building of decision trees from Random Forests in that LightGBM uses an implementation of Gradient Boosting Decision Trees (GBDT), which build decision trees sequentially by exploiting the residual errors of previous decision trees by fitting the negative gradients using Gradient-Based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB) techniques [61], [62]. In contrast, Random Forest trees are determined without iterative refinement and XGBoost builds trees by expanding these level by level [60], [62]. In LightGBM, GOSS prioritises data instances with larger gradients in order to maintain efficiency in downsampling while preserving important information [61]. Meanwhile, EFB organises features into groups, ensuring that only a limited number of features are considered at any given time, without compromising the quality of the model by their contributions [61]. Consequently, LightGBM is well-suited for modelling our large-scale dataset and outperforms the previously discussed XGBoost in terms of computation time and memory usage [61]. LightGBM’s leaf-wise growth strategy is presented in Figure 3.9.

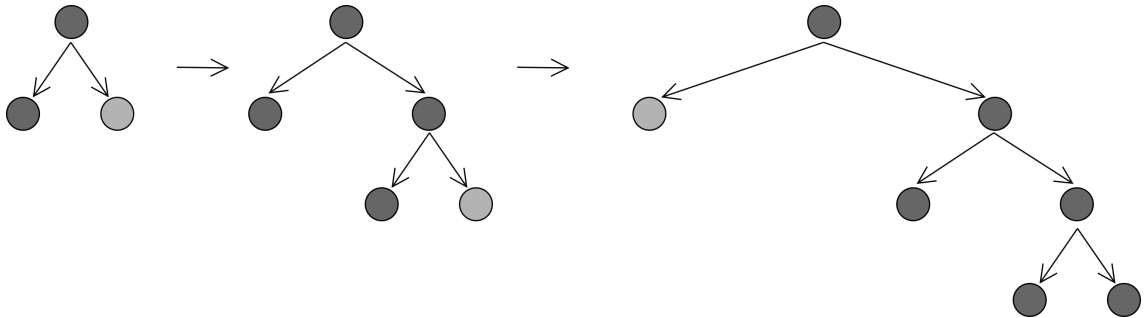


Figure 3.9: LightGBM’s leaf-wise growth strategy visualised. Black indicates terminal nodes, leaves, while grey indicates the node selected to grow next

Easy Ensemble (EE), similarly to Balanced Random Forest, is an ensemble learning method that generates balanced sub-problems but instead uses Adaptive Boosting (AdaBoost) classifiers to solve the given problem [64]. AdaBoost is then a boosting algorithm that combines weak classifiers sequentially to correct errors made by the previous classifiers by changing weights related to the misclassified instances [64]. Easy Ensemble has been developed to handle data imbalance by undersampling subsets of the majority class to train multiple classifiers, thus ensuring that no data is lost as occurs with traditional undersampling methods [64]. These different classifiers, given a subset of the majority class, are given all minority classes to achieve balanced training data, and the previously mentioned AdaBoost algorithm is used for training these classifiers [64]. The results of these classifiers are combined into a single output [64]. A notable distinction between Easy Ensemble and Balanced Random Forest lies in their utilisation of balanced bootstrap samples [63], [64]. While Balanced Random Forest utilises these samples for training decision trees in a random manner, Easy Ensemble utilises them to generate boosted ensembles [64]. Easy Ensemble’s high-level logic is presented in Figure 3.10.

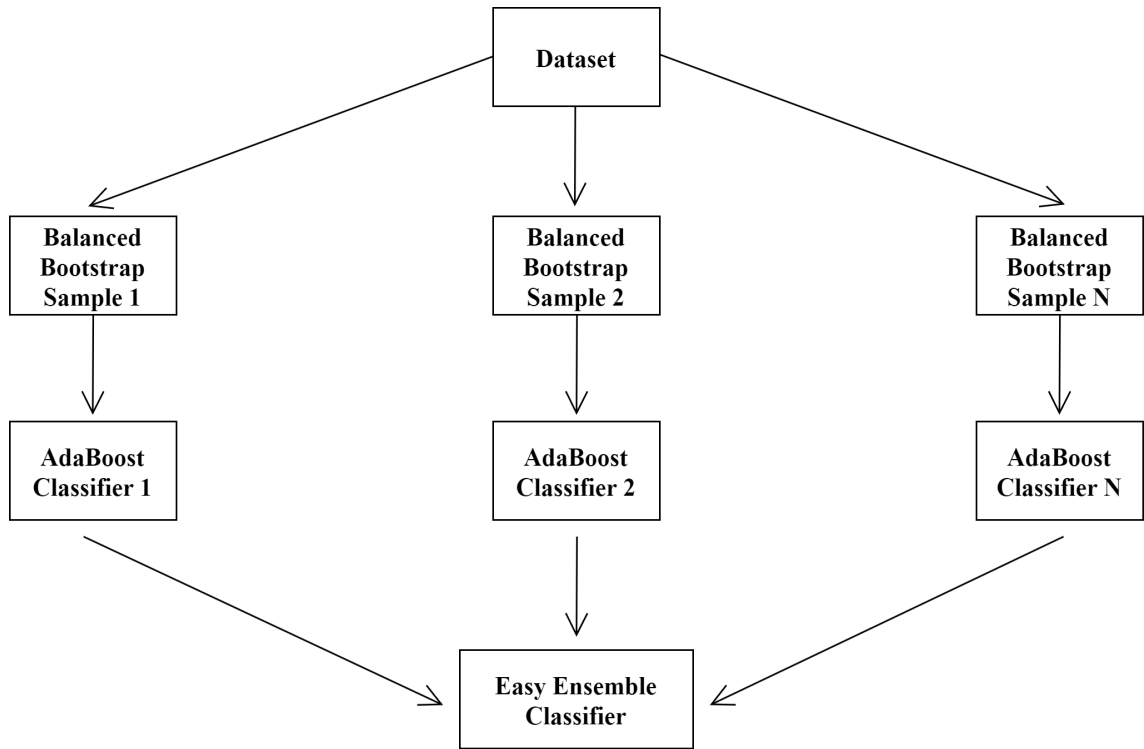


Figure 3.10: AdaBoost with balanced bootstrap samples used to create Easy Ensemble classifier [64]

4 Machine learning approach

This research uses the *Python* programming language (version 3.12) to process data, model and evaluate these models. Libraries such as *SciPy* (v1.14.1), *Scikit-learn* (v1.5.2), *NumPy* (v2.1.2), *pandas* (v2.2.3) and *Matplotlib* (v3.9.2) are also used in support of this research [65]–[69].

In the following subchapters, the discussion will proceed as follows: firstly, the data will be explored, and the processes that were utilised in its handling will be described. This will build on the features and data labels that have been previously mentioned. Subsequently, we proceed to the modelling stage, followed by the presentation of the modelling results.

4.1 Data

The data utilised in this study was obtained from an internal database regarding the aforementioned radiation management system currently employed at TVO. The processing of this data, collected from 2012 onwards at the time of the annual maintenance periods, was initially started through database queries, mainly from the radiation database. Additional data from other relevant databases, such as personnel, skills management and time management databases, were then added to enrich the dataset. The data was retrieved from the database as a CSV file, which was then processed in a separate virtual environment. The initial data and its features are presented in Table 4.1.

Raw Data Variables	
IDENT	123456789
EXT_WORKER	0
ORG_NAME	Teollisuuden Voima Oyj
COURSE_ENTRY	1
COURSE_ADVANCED	1
WORK_START_DATE	24.05.2023 10:51
WORK_END_DATE	24.05.2023 11:15
TIME_USED_MINUTES	24
WORK_DOSE_CODE	200000
PLANT	2
SYSTEM	0
PROJECT	0
WORKER_CLASS	B
RADIATION_DOSE	0
DOSE_ALARM	300
DOSE_RATE_ALARM	250
INFO	OL2 reaktorirakennus, yleiset työt (engl. OL2 reactor building, general work)

Table 4.1: Example data instance (here as randomly generated) after the initial data query. INFO relates to the according WORK_DOSE_CODE

During the preliminary data processing, it was observed that while the data exhibited excellent quality, it did contain some anomalous values, including visits to plants that were inordinately lengthy. Consequently, all data points that exceeded the 13-hour visit limit were removed. The 13-hour limit was selected as this is the duration at which the electronic dosimeter issues an alert regarding the visit length.

The data was then scanned for duplicates and these were removed as a precaution. The total number of data points removed was a few thousand. In addition to these anomalies, the data lacked personnel classification information for approximately 100 plant visits. This was resolved by calculating the doses of the personnel in question and comparing it with the legal regulations, thereby providing a rationale for the categorisation of the personnel who made the visit as either category A or B. Category A personnel can receive a maximum effective dose of 20 mSv per year and category B personnel can receive a maximum effective dose of 6 mSv per year [6], [16]. In addition to these preprocessing steps, the timestamps of the plant visits were tested to ensure that the exit time could not be prior to the entry time. No such erroneous visits were found.

Upon closer inspection of the preliminary filtered data, it was confirmed that the radiation dose data deviated from a normal distribution, with a bias towards lower doses. This is because most occupations do not involve high levels of exposure to ionising radiation. This skewness can be observed in Figure 4.6, which shows the classed radiation doses after data preparation.

The data were further processed, involving the transformation of variables into either categorical or continuous types, among the variables, only 'dose alarm' and 'radiation dose' were defined as continuous, while the remainder were set as categorical. The rationale behind setting the 'dose alarm' as continuous is due to its dynamic nature, whereby it is adjusted based on the characteristics of each individual's visit and the specific dose threshold, as previously outlined in this study. The radiation dose was used to construct classes for the machine learning classification task, as discussed later. After the type conversions, we proceeded to outlier detection, which is the removal of data points that are significantly different from the rest of the data [70]. This was done because the aim of the study was not to model anomalies, but specifically the visits and their respective typical doses. We used the Interquartile

Range (IQR) proximity rule in this task because the radiation doses were not normally distributed [70]. IQR is defined as the range of values between the first and third quartiles of a data set. That is to imply that if a data point differs from the lower limit of the 25th quantile by -1.5 times the IQR and from the upper limit of the 75th quantile by +1.5 times the IQR, then the data point is classified as an outlier [70]. IQR can be calculated as follows:

$$x < Q_1 - 1.5 \times (Q_3 - Q_1) \quad \text{or} \quad x > Q_3 + 1.5 \times (Q_3 - Q_1)$$

The identification of these outliers using the IQR method resulted in the removal of tens of thousands of instances of data. The majority of these deleted instances, or site visits per WDC, were for the generic WDCs: 100000 and 200000. These WDCs accounted for more than half of the total number of instances that were deleted. This suggests that the generic codes are being misused because of their ease of use. In this case, the personnel does not need to remember or know the exact WDC required for the work, or that there is no specific WDC in use that would be needed to exclude these outliers from the rest. Removing these outliers significantly improved the performance of the machine learning models, which we discuss further in Chapter 4.2. Figure 4.1 shows the percentage distribution of outlier counts per year for both sites. This figure also demonstrates how the type of annual maintenance (fuel change and maintenance and fuel change only) varies from year to year and affects the outlier amounts.

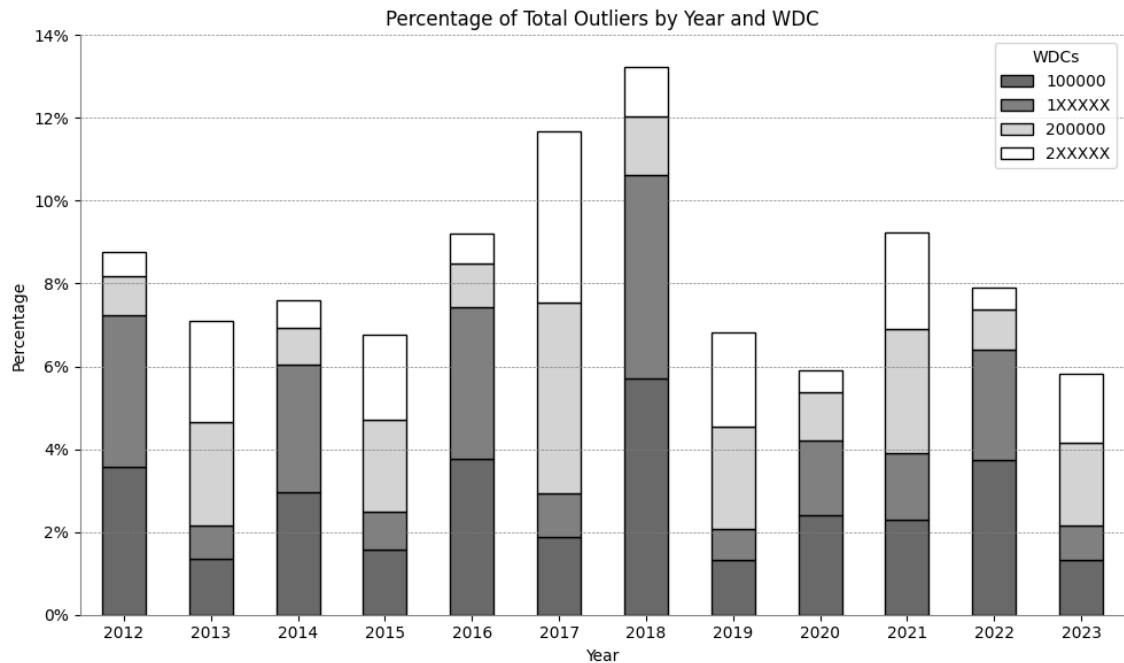


Figure 4.1: Outliers yearly from the total percentage of outliers. Grouped by general WDCs and rest are bundled together by plant (OL1 or OL2)

The raw data and the visualisations derived from them are not presented in this study, and all the Statistics, Figures and Tables from this point onwards have been derived from preprocessed data.

Once the outliers have been identified and removed, the data can be examined in greater detail for the published part of this study. Following the removal of outliers, the size of the dataset remains in the hundreds of thousands. Firstly, the objective is to visualise the trend between the time and dose. As demonstrated in Figure 4.2, an increase in visit time does not appear to have a significant impact on the dosage, with the high-dose work being completed promptly. The heatmap representation further reveals that the data is heavily skewed towards lower doses, which are also more proportional to the time spent. The absence of prolonged high-dose visits is evident in the heatmap, as there are no data points to model such visits.

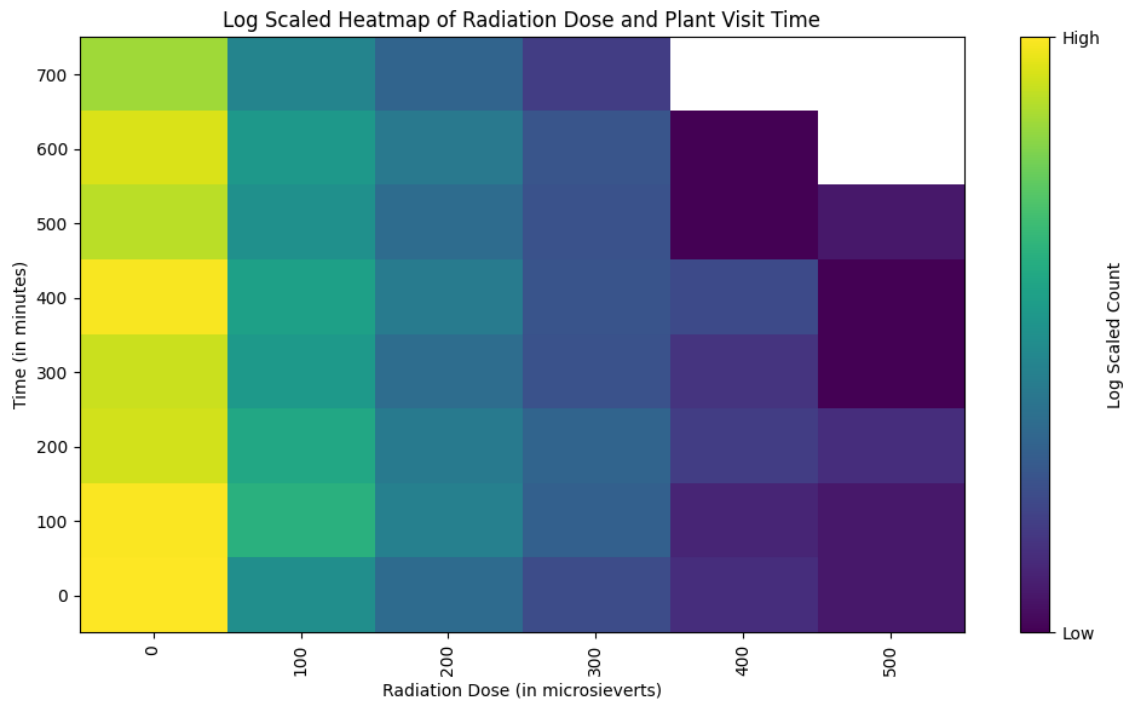


Figure 4.2: Heatmap presentation of visit length and the according radiation dose in a logarithmic scale

At TVO, personnel undergo training in radiation work through two distinct courses: firstly, there is a compulsory entry-level course, and secondly, there is an advanced course, which is not compulsory but is attended when necessary. Figure 4.3 illustrates the compulsory and advanced courses in their respective boxplots. However, the interpretation of these plots is complicated by the fact that the doses received by both groups are very close to zero. To provide a more meaningful interpretation, logarithmically scaled plots were created, which revealed that personnel who completed the advanced course received equivalent low-level radiation doses compared to the personnel who only completed the entry-level course, yet at higher doses, those who completed the advanced course received fewer doses. This is further supported by the statistical analysis which showed that the mean dose received by personnel who had only completed the entry-level course was approximately $2 \mu\text{Sv}$ higher than the mean dose received by personnel who had also completed the advanced course.

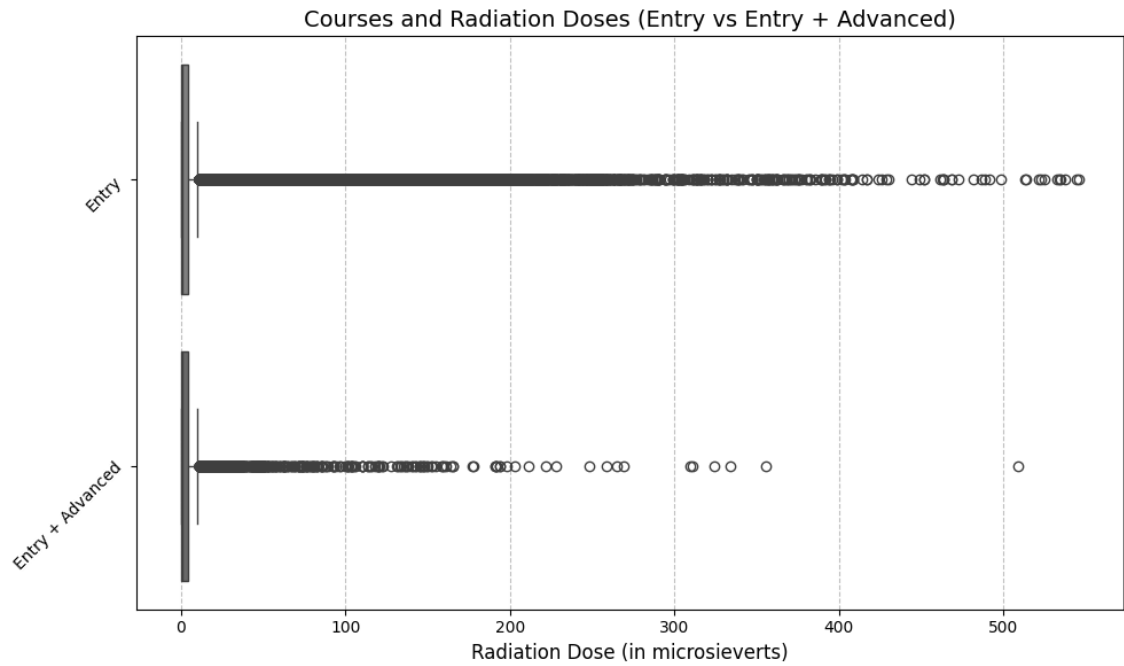


Figure 4.3: Boxplot of radiation doses and courses completed by personnel. For data that was normally undistributed and relational, the Wilcoxon signed-rank test was used to show whether there were differences in doses. A p -value of 0.0 indicates a significant difference between radiation doses between the groups

We then sought to determine whether any of the work dose codes exhibited a high degree of significance with respect to doses. As illustrated in Figure 4.4, it was evident that the general codes (X10000) for the containment building and the work with the actuators for the control rods (WDC X22100) had the highest accumulated doses since 2012, particularly for the OL1 plant. The generic codes were expected to become over-represented in this statistic, as can be seen particularly for the OL2 plant.

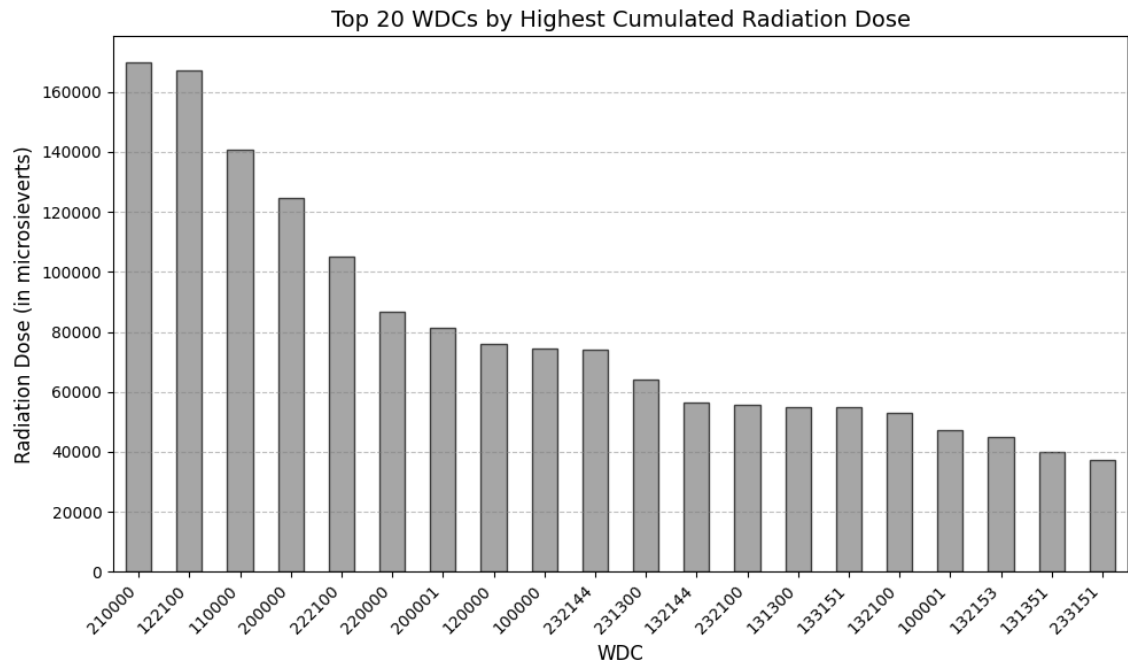


Figure 4.4: Accumulated doses for WDCs visualised since 2012 annual maintenances

However, if we separate the system from the WDCs, we can see from Figure 4.5 that, in common between OL1 and OL2 plants, the cooling system of the shutdown reactor has produced the highest doses (321), also more than the general codes. We can see that systems 221, 331, 313 and 200 in particular produce the highest doses, if we exclude 321 mentioned before, and the general systems 100 and 0. From the general codes, it is not possible to deduce more specific reasons for the visits, which is challenging for this study and for doses gathered, as interpretability is reduced. Therefore, an analysis of the variation in such doses was conducted, which revealed a significant amount, suggesting potential misuse of the codes.

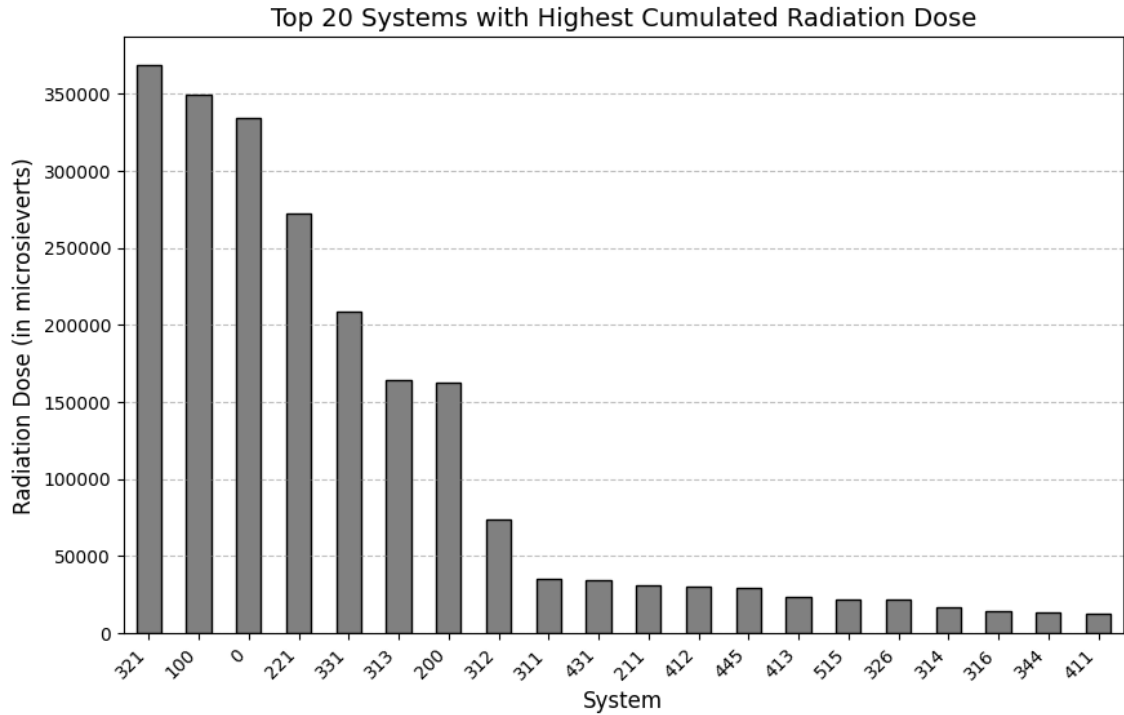


Figure 4.5: Accumulated doses for OL1 and OL2 plant systems since 2012 annual maintenances. System is parsed from the WDCs, for example 132100 would mean OL1 321-system, which is the cooling system of the shutdown reactor

After exploring the data, we wanted to use feature generation to add informativeness through domain-specific knowledge [58]. Therefore, we decomposed parts of WDCs into their own features as we already demonstrated earlier. We also added a variable for the type of annual maintenance (fuel change and maintenance or fuel change only) and added a feature based on the time spent on the visit to indicate whether the visit was short, medium or long (less than 2h, less than or more than 8h). We also removed the variables related to organisations, WDC info, dates and IDs because we thought that it would be dangerous to teach the model to predict doses based on these variables, let alone by date or ID. We then created a discrete class distribution from the dose, removing the continuous radiation dose from the data to prevent data leakage. This continuous value (expressed in microsieverts) was classified into four distinct classes, labels 0 to 3, with their respective intervals specified in Table 4.2. The classification intervals were determined through a com-

bination of domain expertise and random search, with the objective of maximizing the informative value of the labels. This approach entailed randomly shifting the interval boundaries while maintaining proximity to the initially domain-knowledge described intervals.

Class label	Class interval*	Ratio**
0	[0, 5)	53.8
1	[5, 25)	9.31
2	[25, 75)	2.45
3	[75, 546]	1.0

Table 4.2: Class labels and their respected intervals and ratios from 0 to max dose of 546 μSv

* = expressed in microsieverts (μSv)

** = relatively scaled each class count by normalizing it against the smallest class size to highlight skewness after data preparation

Based on the ratios presented in Table 4.2, it is clear that the distribution of radiation doses is highly skewed. The mean dose is substantially higher than the median and the standard deviation of 21.29 μSv further indicates variability in the data, with doses ranging from a minimum of 0 μSv to a maximum of 546 μSv . The 75th percentile of the doses falls below the 4 μSv threshold, suggesting that higher values have a significant impact on the overall distribution.

Subsequently, given the skewed and time-series related nature of the temporal data, we split these into the training, validation and testing sets (60/20/20) as mentioned earlier. This split was done so that the oldest data instances ended up in the training data and the most recent ones in the testing data, to avoid data leakage from the past to the future and to use historical data to present how doses have evolved through time [71]. The training data included data points from 2012 to 2018, validation from 2018 to 2021 and test data from 2021 to 2023. This analysis would be seasonal if we used all the data available, but we removed this aspect by

modelling only the periods associated with annual maintenance [72]. Skewed dose classes after splitting are illustrated in Figure 4.6.

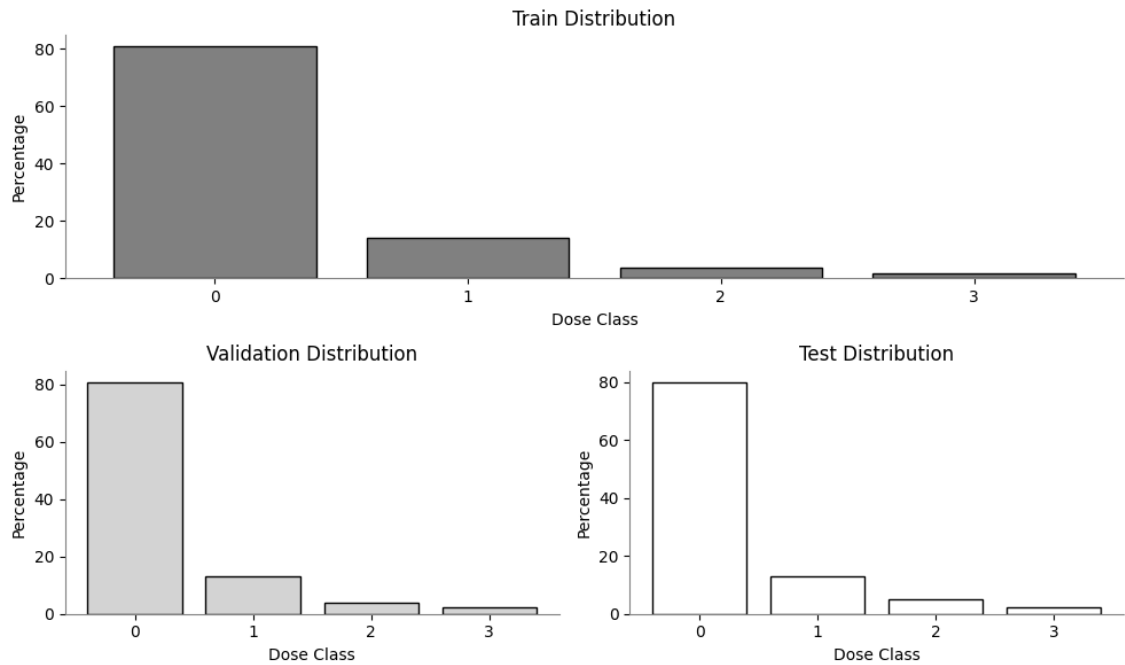


Figure 4.6: Skewed radiation dose classes after train/validation/test split

As a result of these data processing steps we have features that are delineated in Table 4.3. Categorical variables were one-hot encoded before assigning the features to be trained. Normalisation of continuous variables was not performed due to the models employed. As a result, we ended up with a few hundred features.

Variable	Description	Data Type
External	Boolean indicating if the personnel is external	Categorical
Entry-course	Boolean indicating if the course is completed	Categorical
Advanced-course	Boolean indicating if the course is completed	Categorical
Time in minutes	Site visit length	Continuous
Work dose code	Code for entry to controlled area	Categorical
Plant	Plant that is entered	Categorical
System	System inside the plant	Categorical
Project	Project for the specified system	Categorical
Personnel class	Radiation work classification	Categorical
Dose alarm	Dose alarm used for the entry	Continuous
Dose rate alarm	Dose rate alarm used for the entry	Categorical
Time category	Low, medium or long visit time	Categorical
Outage	Type of the annual maintenance	Categorical

Table 4.3: Variables feature engineered from the raw data

4.2 Applying machine learning

Moving forward, we will utilise the data that has been preprocessed and the models that have been presented in order to proceed with the machine learning. The models presented in Table 3.1 were trained using training data with different parameters. These models with different parameters were then evaluated using validation data, with the test set being kept truly separate from the training and validation phases throughout this process. Just under 500 different parameter combinations were validated between the models, and the best combination per model was selected based on the results obtained. These best-performing models with the respected

parameters were then tested using test data. The models and their validated parameters are presented in Table 4.4.

Model	Hyperparameters
Random Forest	max_depth: 3, 5, 7 n_estimators: 800, 1200, 1400, 1600 max_features: 'sqrt', None min_samples_leaf: 1, 3, 10, 15
Balanced Random Forest	max_depth: 3, 5, 7 n_estimators: 800, 1200, 1400, 1600 max_features: 'sqrt', None min_samples_leaf: 1, 3, 10, 15
XGBoost	n_estimators: 800, 1200, 1400, 1600 learning_rate: 0.01, 0.05, 0.1 max_depth: 3, 5, 7 subsample: 0.8, 1.0 min_child_weight: 1, 2, 5
LightGBM	n_estimators: 800, 1200, 1400, 1600 learning_rate: 0.01, 0.05, 0.1 bagging_fraction: 0.8, 1.0 feature_fraction: 0.8, 1.0
Easy Ensemble	n_estimators: 10, 20, 30, 50, 100

Table 4.4: Hyperparameters for the different models validated in this study

While validating these parameters, it was found that the imbalance of the classes, and in particular the limited number of data instances in the minority classes, resulted in challenges in predicting them, as would be expected. The problem was addressed by rebalancing the minority classes inversely proportional to their frequencies, which

resulted in improved validation results. These weights were imposed on all models except Easy Ensemble and Balanced Random Forest, as their internal structures are capable of handling the balancing [63], [64]. For the remaining models, internal balancing was not used because they did not produce significantly different results. Instead, these models were given weighted classes with the data.

An attempt was made with a data centered approach to resolve the said data imbalance by using naive down- and oversampling methods for the training data, such as randomisation, but these did not yield better results [73]. Rather, these worsened the results because downsampling removes informativeness by removing data points and oversampling produces noise to the dataset and can lead to overfitting [73]. More advanced approaches also exist, but these were not pursued with this study.

In the next Chapter 4.3 we will address the metrics used in this study to find the most optimal parameters obtained through validation. The same metrics are also used for testing that is achieved by combining the training and validation data and re-training the models using this data. The models will then be tested against the test data and analysed to identify the areas of success and the areas that require further development in Chapter 5.

4.3 Evaluation

In this study, Area Under the Receiver-Operating Characteristic (ROC AUC) and Weighted F1 scores were selected as the validation and testing metrics. The definitions of these metrics can be found in Table 4.5. In addition to these metrics, the ROC and Precision-Recall curves (PR-curves) were analysed as well as Confusion Matrices, alongside hyperparameter validation and final testing.

Metric	Definition
ROC AUC Score	$\text{TPR (True Positive Rate)} = \frac{\text{TP}}{\text{TP} + \text{FN}}$ $\text{FPR (False Positive Rate)} = \frac{\text{FP}}{\text{FP} + \text{TN}}$ $\text{ROC AUC} = \int_0^1 \text{TPR}(x) d(\text{FPR}(x))$ <p>where TP is True Positives, FP is False Positives, TN is True Negatives and FN is False Negatives</p>
Weighted F1 Score	$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$ $\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$ $\text{F1}_{\text{weighted}} = \sum_{i=1}^N w_i \cdot 2 \frac{\text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}$ <p>where $w_i = \frac{n_i}{N}$, n_i is the number of instances in class i, and N is the total number of instances</p>

Table 4.5: ROC AUC and Weighted F1 metrics and their respective definitions [9], [74]

The metrics selected were based on the premise that the dataset displayed significant imbalance, a situation in which traditional accuracy metrics may prove to be misleading due to the dominance of the majority classes [9], [75]. ROC AUC metric was selected as it offers a balance between the True Positive Rate and the False Positive Rate (the larger the ROC AUC the better, 0.5 signifies random and 1.0 perfect chance) [11], [74], [75]. In imbalanced datasets ROC AUC remains

informative due to its independence from the changes in the class distribution [74], [75]. ROC-curves, on the other hand, can visualise the performance of classifiers regardless of class imbalance [75]. AUC is therefore the probability that the classifier will rank a randomly selected positive instance higher than a randomly selected negative instance [74], [75]. Conversely, the F1 score considers both Precision and Recall [9]. Precision ensures that the model does not misclassify negative instances as positive, while Recall ensures that the model detects positive instances as many as possible, meaning in that high Precision means low FPR and high Recall low FNR (False Negative Rate) [9]. The Weighted F1 score functions similarly but adds the weights for each class based on the number of true instances in that class, thus reflecting performance more accurately for minority classes. Table 4.6 shows the validation results for the best hyperparameters selected by ROC AUC score.

Model and Results	Selected Hyperparams
Random Forest Results: ROC AUC: 0.8434 F1 Weighted: 0.7607	max_depth: 7 n_estimators: 1600 max_features: 'sqrt' min_samples_leaf: 1
Balanced Random Forest Results: ROC AUC: 0.8447 F1 Weighted: 0.7603	max_depth: 7 n_estimators: 800 max_features: 'sqrt' min_samples_leaf: 1
XGBoost Results: ROC AUC: 0.8997 F1 Weighted: 0.7708	n_estimators: 1200 learning_rate: 0.05 max_depth: 7 subsample: 0.8 min_child_weight: 5
LightGBM Results: ROC AUC: 0.9009 F1 Weighted: 0.7674	n_estimators: 1600 learning_rate: 0.01 bagging_fraction: 0.8 feature_fraction: 0.8
Easy Ensemble Results: ROC AUC: 0.7706 F1 Weighted: 0.6458	n_estimators: 10

Table 4.6: Validation results for the models used, including hyperparameters and performance metrics ROC AUC and Weighted F1 score

From the validation results in Table 4.6, it is clear that LightGBM produces the best results, which is also reflected in the ROC AUC score for the minority classes. There is also not much difference between Balanced and traditional Random Forest

models, except for the predictions of class 3. However, it is already evident that the models are struggling to accurately predict the minority classes. The best validation result obtained with LightGBM was most affected by the change in learning rate, which is visualised in Figure 4.7 using different metrics.

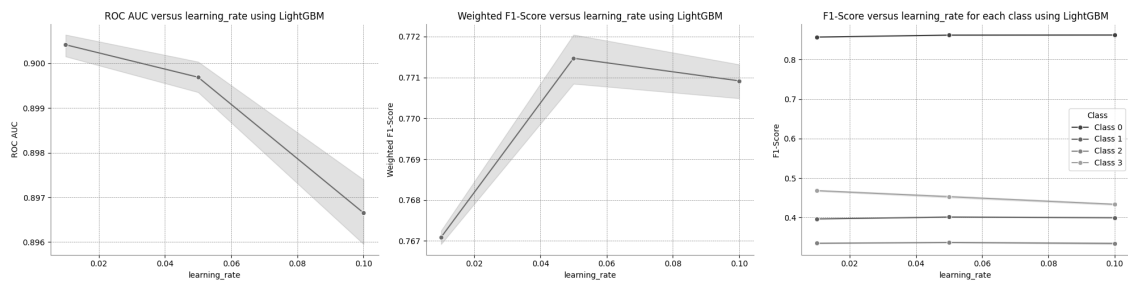


Figure 4.7: LightGBM learning rate parameter and its effect on respected metrics

According to the validation results, in the tree-based models, a trend was identified that suggests a positive relationship between tree depth and the ROC AUC score, along with a variable impact on the Weighted F1 score. However, the number of trees alone exhibited minimal influence on model performance, with the exception of XGBoost. Notably, LightGBM demonstrated a decline in performance with increasing tree depth when this parameter was considered independently. Among the models examined, Easy Ensemble exhibited significantly poorer performance compared to the other models, as reflected also in its weaker performance in majority class classification.

5 Results

After acquiring the test results, it can be determined, as in the validation, that the LightGBM model demonstrates the most optimal performance in terms of our classification task, despite encountering challenges. These challenges relate to the model's inability to effectively differentiate between minority classes. Consequently, the majority class ($0 - 5 \mu\text{Sv}$) can be distinguished from the other classes with a high degree of confidence, while the higher doses (above $75 \mu\text{Sv}$) can be distinguished with a medium degree. However, the intermediate classes 1 and 2 remain challenging to distinguish with the available data. This can also be seen in the PR-curve behaviour shown in Figure 5.1. PR-curve visualises Precision-Recall tradeoff at different thresholds [75]. The models were also found to be considerably more effective after combining training and validation data prior to testing against test data.

With the exception of LightGBM, the interpretation of other models via the PR-curves and Weighted F1 results has been shown to be equally effective in the classification of classes 1 and 2. Random Forest (also balanced) performed better in the classification of class 3 at the expense of classification of the class 0.

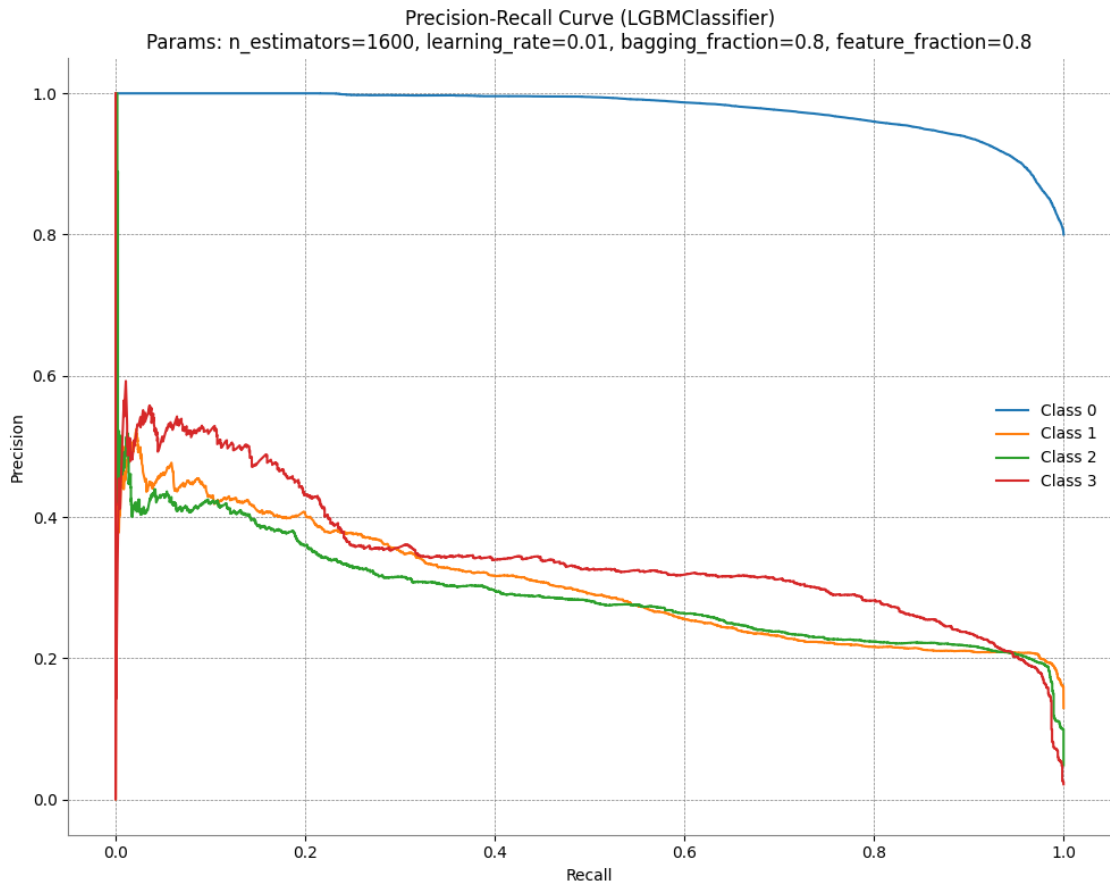


Figure 5.1: LightGBM Precision-Recall curve

The models were also validated and tested without general WDCs due to the noise they contain, yet this did not result in a positive impact on the models performances. However, their incorporation did result in an increase in the mean ROC AUC and Weighted F1 score by approximately 20% on average. Furthermore, a scenario was created in which outliers were not removed, which resulted in a substantial decline in performance, both with and without the generic WDCs.

In comparison to other models, LightGBM and XGBoost achieved higher correct classification rates for instances within classes 1 and 2. However, their performance was suboptimal for class 3 predictions. The Confusion Matrix of LightGBM is presented in Figure 5.2. In contrast, Random Forests showed a 10% decrease in the correct classification rates for classes 1 and 2, yet exhibited almost a 15% increase

in class 3 predictions. For the class 0 classification, Random Forest demonstrated a comparable performance to LightGBM and XGBoost, as illustrated by the Confusion Matrix in Figure 5.3. However, no significant differences were observed between the balanced and traditional Random Forest models. The classification performance achieved by the Balanced Random Forest with classes 0 and 1 was marginally better than that of the Random Forest.

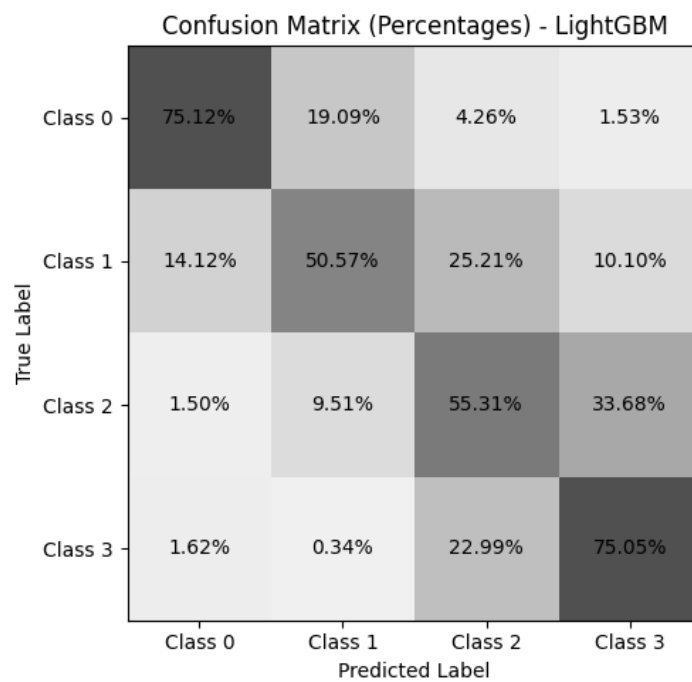


Figure 5.2: LightGBM Confusion Matrix

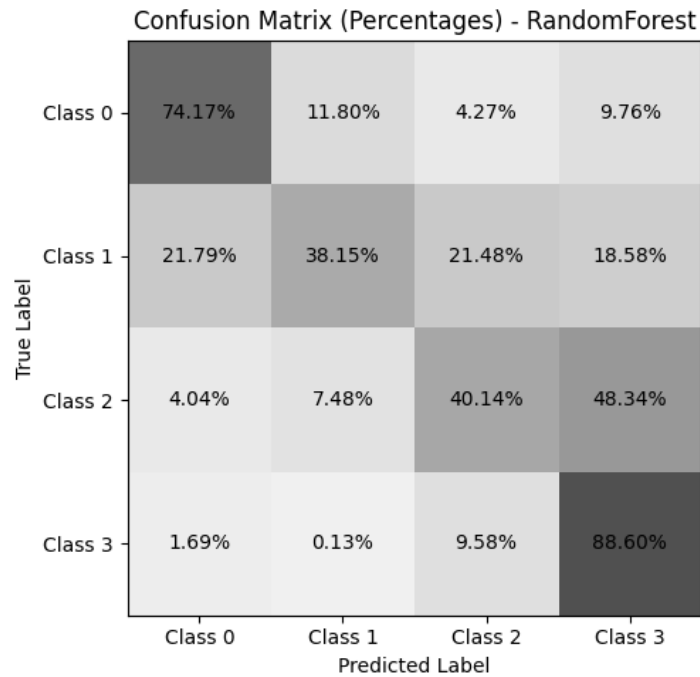


Figure 5.3: Random Forest Confusion Matrix

However, when solely interpreting Confusion Matrices, it can be observed that only classes 0 and 3 achieve usable results while the results of classes 1 and 2 are comparable to a random chance. It can be observed that the models predict larger classes on average, which suggests that the models are not able to distinguish between classes, especially for the higher classes.

As is evident, interpreting numerous classifications poses a challenge in managing the entirety of the output space. When considering a classification of only four classes, the Confusion Matrix transforms into an $n \times n$ matrix, comprising $n^2 - n$ potential error classifications [75]. In our case, $4^2 - 4 = 12$. This approach, however, is not without its limitations. The prediction of classes in this case is based on *argmax*, where the highest probability between the classes is taken, a method that may not always be correct, especially if the probabilities are close to each other as a result of imbalance. This is a weakness of the Confusion Matrices, as it is a threshold-specific metric. To address these limitations and the resulting complexity, we created ROC

curves for each class using a One-vs-the-Rest (OvR) approach, where we designate one class as positive and the others as negative [75]. This approach was also used to estimate the AUC score [75]. Due to the imbalance, their average was estimated by micro-averaging, a method that treats each class as binary and aggregates the average of the contribution of each class. Therefore, the ROC AUC curves are used to examine the model's ability to distinguish between classes [75]. Through the comparison of all of the metrics, it is possible to address the model's capacity to differentiate between classes and translate that ability into predictions at the decision level. The ROC AUC curves for LightGBM and Random Forest are presented in Figures 5.4 and 5.5, respectively.

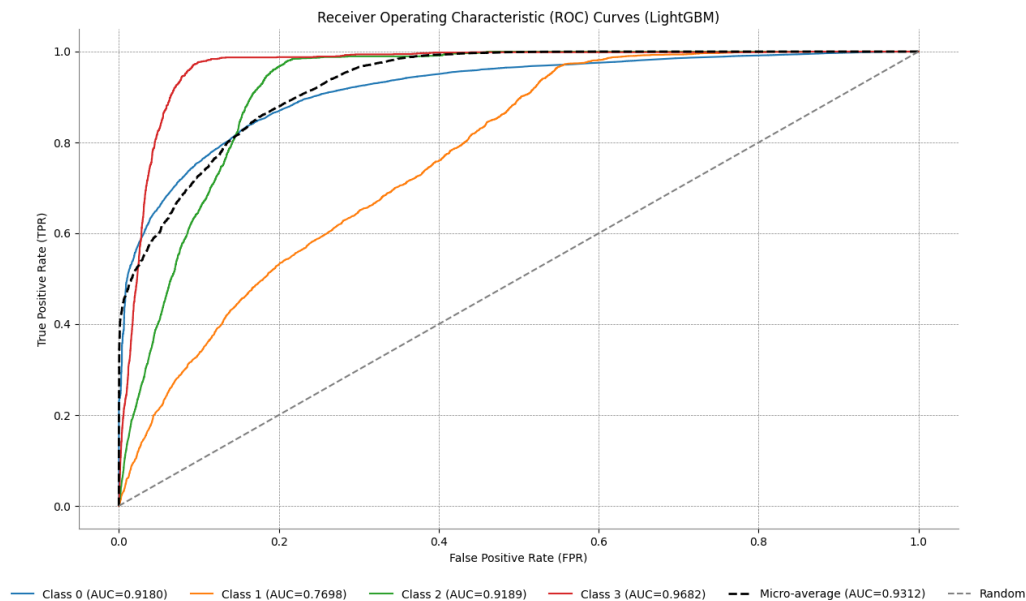


Figure 5.4: LightGBM ROC AUC curves, AUC and micro-average AUC scores

As can be seen, LightGBM achieves higher AUC values at all thresholds, which is also reflected in the micro-average AUC value. In particular, LightGBM improves significantly for class 0 and class 2. For both models, class 3 is the easiest to distinguish. Similar behaviour can be observed for the other models considered, among which Easy Ensemble stands out significantly with its weaker ranking capabilities.

Furthermore, when Confusion Matrices are taken into account, LightGBM achieves a better balance across classes, while Random Forest relies heavily on its strong performance for class 3. Based on these metrics, including Weighted F1 scores, it can be concluded that LightGBM has more consistent performance across classes, while Random Forest has class-specific strengths.

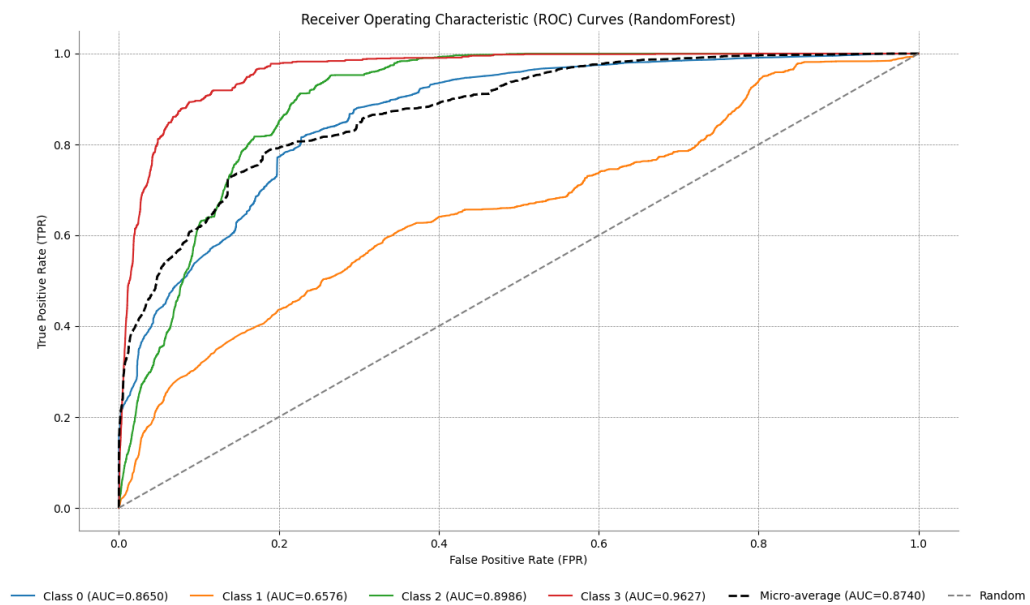


Figure 5.5: Random Forest ROC AUC curves, AUC and micro-average AUC scores

However, these findings suggest that by restructuring the experimental approach as a binary task, with an exclusive focus on predicting extreme values, the performance of the models could be significantly improved. While the current approach does not yield optimal results, there is nonetheless considerable value in the prediction of extreme outcomes.

In the follow-up analysis, the feature importances are assessed to understand which features the models prioritise to achieve these results. As illustrated in Figure 5.6, the feature importances of LightGBM reveal its significant reliance on time spent, which is reasonable given that longer visit times result in a lower average dose, while shorter visit times potentially lead to a higher dose on average (high-

dose work is completed promptly). It is noteworthy that the dose limit and worker classification hold considerable importance, with the high-dose systems also impacting the predictive outcomes as we saw with the data analysis. A particularly noteworthy feature is the presence of an OL2 plant, which may imply that the doses at the OL2 deviate from those at the OL1, despite both being identical. The presence of external personnel may indicate that such contractors could be exposed to higher doses than regular personnel due that their specialized knowledge is often required for specific work tasks which can result in higher doses.

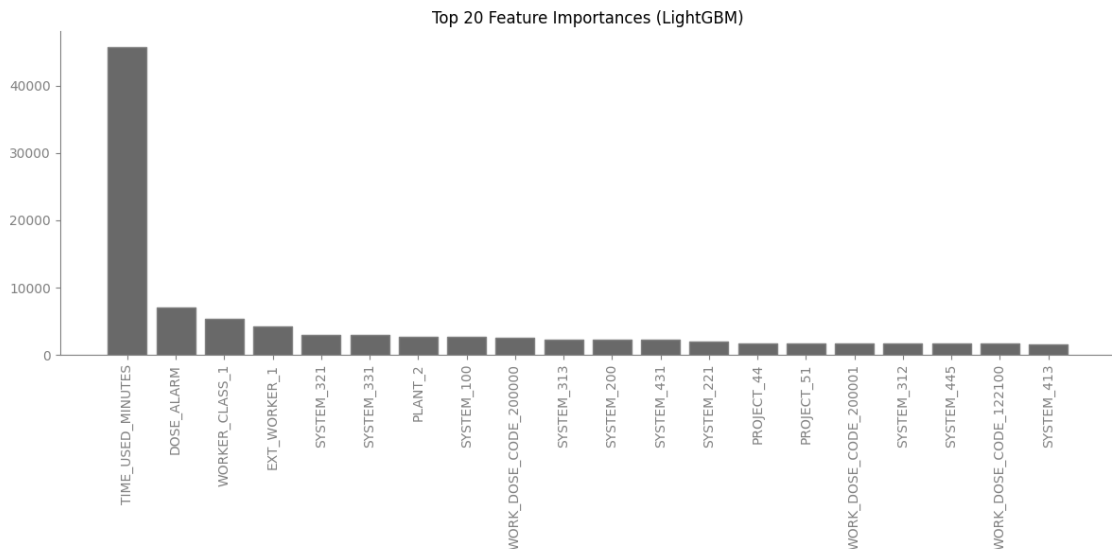


Figure 5.6: LightGBM feature importances

However, Random Forest places importance on different features compared to LightGBM. While both models prioritize the time spent on a visit as the most significant feature, Random Forest distributes its reliance more evenly across other features, as shown in Figure 5.7. Nevertheless, LightGBM and Random Forest share the same systems as most important features, but this was also the case with other models. Balanced Random Forest favours almost identical features compared to traditional Random Forest, a result that was to be expected. In contrast, XGBoost places significant emphasis on workload code 100001 and project

01, although this is distributed more uniformly across all features when compared with LightGBM. However, LightGBM consistently achieves optimal results across all metrics, suggesting that time is the most significant predictor and the personnel dose limit and category are of considerable importance. The presence of the OL2 in the most important features necessitates further observation, as it is an outlier that other models fail to identify.

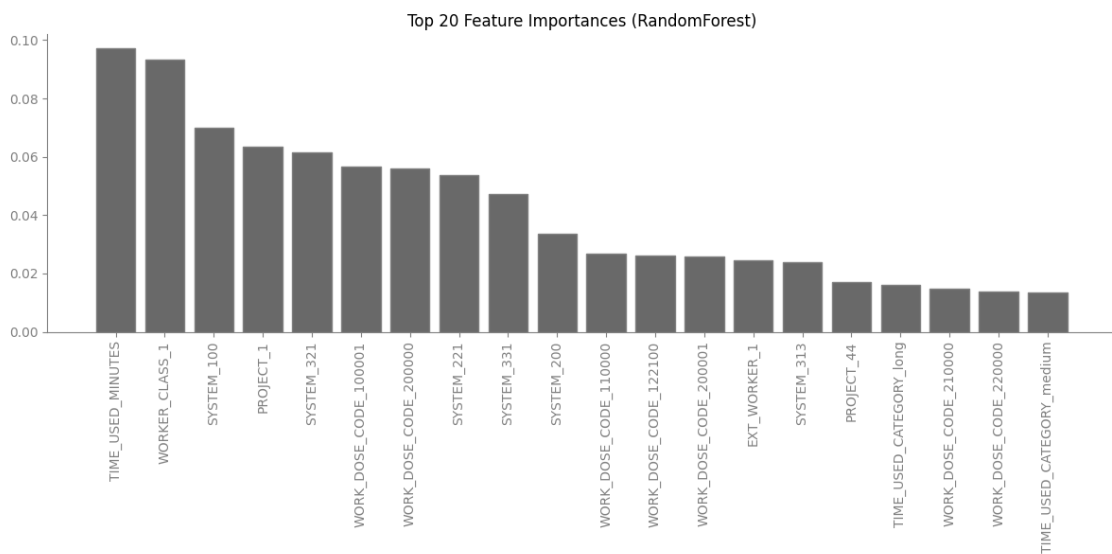


Figure 5.7: Random Forest feature importances

In summary, LightGBM achieves the best ROC AUC and micro-averaged scores, while XGBoost achieves the best Weighted F1 score. The Random Forests achieved comparable scores, while Easy Ensemble performed significantly worse in all metrics. LightGBM achieves the most optimal results, demonstrating that it is the most efficient model in this comparison. However, it falls short of delivering the level of performance required for the classification task at hand. A summary of the results can be found in Table 5.1. From the point of view of achieving the objective of this study, we can conclude that this could not be achieved with the available data.

Model	Test Results
Random Forest	ROC AUC: 0.8460, Weighted F1: 0.7325
Balanced Random Forest	ROC AUC: 0.8454, Weighted F1: 0.73088
XGBoost	ROC AUC: 0.8931, Weighted F1: 0.7532
LightGBM (Most optimal)	ROC AUC: 0.8938, Weighted F1: 0.7519
Easy Ensemble Classifier	ROC AUC: 0.7705, Weighted F1: 0.6168

Table 5.1: Overall test results with ROC AUC and Weighted F1 scores

5.1 Evaluation analysis

In addition to the prior data extraction in Chapter 4.1, an attempt was made to integrate the work order system data with the radiation data. However, this proved to be a particularly challenging task, which would have enabled a significantly more accurate assessment of the scope of the visits. The system associated with each site visit is known from the radiation database, however, integrating data from the work order system would have provided more detailed insights into the specific tasks performed to the component level during each site visit. Unfortunately, the attempt to link the data by personnel ID and the time of the visit and the work proved to be unsuccessful. This was due to the fact that a single personnel may have had multiple work activities in progress simultaneously, which resulted in the inability to determine the precise scope of the site visit. This would have resulted in the generation of numerous false positives, amounting to several million.

This serves as a reminder that the quality and informativeness of data is not always sufficient for machine learning. In this particular instance, it can be concluded that the radiation data lacks the necessary level of detail to generate reliable and informative results. The lack of integration between different databases, which would have been essential in this case, highlights an opportunity for future improvements to enhance the informativeness of the data.

In addition to the issue of data informativeness, it may also be necessary to consider more advanced methods for data over and downsampling, especially if the classification task is to be made more precise by reducing the size of the class intervals. At present, this is a highly challenging task, as obtaining data for all classes and splits may not be possible due to time dependencies in the data. This is due to the fact that, at higher doses, the necessary data points simply do not exist.

The situation is made even more complicated by the fact that the model in question is meant to be used in a field that's critical to safety and health. This makes the performance expectations for the model extremely high, and unfortunately, these requirements weren't met.

However, it is important to acknowledge that this was the first experimental setting for such a problem, which also provided valuable insights into the limitations and possibilities of applying machine learning to similar tasks or in a broader context. Moreover, the limitations and findings of this study offer an important source of insight for similar tasks at TVO.

5.2 Operational use

In practice, however, the implementation of machine learning and related approaches does not function on a create and forget basis, yet rather demands continuous development and monitoring. Figure 5.8 below illustrates the life cycle of an ML-enabled software system. It is important to note that these stages do not necessarily

follow a waterfall pattern, as they can also occur in different orders or in parallel [76]. The figure then illustrates the essential steps required to deploy and upkeep the ML solution in production.

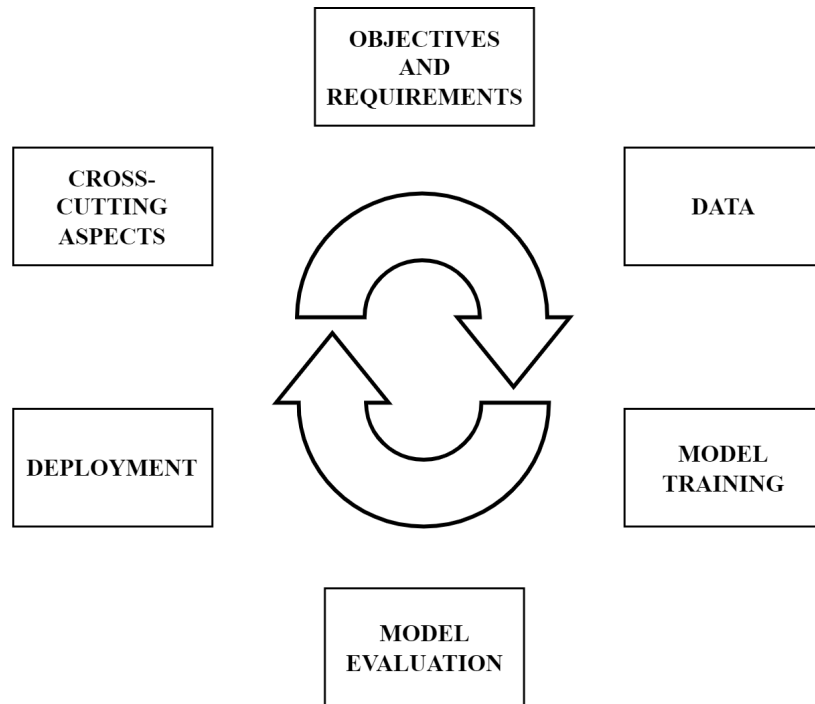


Figure 5.8: Lifecycle of ML-enabled software system [76], [77]

In our case, if a similar, however, better performing machine learning model were to be put into production use, the readiness and suitability of the data must first be assured, which requires, as in this case, domain expertise [78]. It is also important to ensure that even if the data is suitable, that this corresponds to operational realities, for example, whether we can always use certain features in the prediction [78].

Nevertheless, machine learning, its interpretability, the data and its processing represent only a fraction of the entire system that requires consideration. At this point in the research, we will focus on the steps following the model evaluation in its broader context. It has previously been demonstrated that the quality of the modelling itself is dependent on the data and its respective issues and as such, these will not be revisited. However, it is still important to acknowledge the existence of

additional limiting factors associated with modelling, in addition to what we have already described in this study, including computational resources, complexity, and regularisation problems [76].

In this context, model deployment includes the integration, monitoring and maintenance of the models [76], [77]. The next phase of cross-cutting concerns brings broader ethical, legal, trust and security issues into the picture [76]. These will be briefly discussed below.

Integration means creating the infrastructure and interfaces for the model, but also the workflow itself, where it is important to involve the people who created the model with those who maintain the runtime environment, which helps to deliver benefits in terms of quality and speed of product delivery [76].

This brings us to the next stages of monitoring and maintenance, where we can ensure that the model is actually functioning as it should and is not deviating from normal and expected behaviour [76]. This means that, over time, models may deviate as data no longer aligns with their original training. [76]. These are so-called concept drifts, where, for example, an external event affects the input data over time [76]. This enables continuous learning, allowing retraining as the model's shortcomings emerge [76], [77].

In the process of analysing the cross-cutting aspects of the operational use, it becomes visible that ethical concerns result from biases that are embedded in the training data, which in turn can lead to unintended discrimination [76]. Conversely, legal challenges stem from the necessity for regulatory frameworks, such as the General Data Protection Regulation (GDPR), which are designed to protect sensitive data, but often lag behind the rapid pace of AI development [76]. In the context of building trust with end-users, it is important to emphasise transparency, clear communication, and the interpretability of models that are tailored to specific use cases [76], [79], [80]. Also the use of applied black-box models should also be

considered, as these are very difficult, if not impossible, to interpret [79]. Security concerns, such as adversarial attacks including data poisoning, model stealing and model inversion, must be addressed when deploying such models [76].

In the case of TVO, this signifies that in the future, the implementation of a comparable AI solutions will require the dedicated infrastructure and computational resources. This emphasises the need for not only proficient personnel in infrastructure administration but also specialised knowledge in data analysis and the field of AI. It is also important to address the role of model monitoring in ensuring the reliability of future predictions, a matter with direct consequences for the quality of the predictions. The use of models, particularly deep learning models, calls for cooperation with information security, even if the models are run on the intranet due to the previously mentioned issues. However, it is apparent that the most significant initial step is the recruitment of skilled personnel to carry out similar projects in the future.

6 Conclusions

The objective of this study was to address radiation and its effects in the nuclear power plant environment according to current standards and ultimately to predict radiation doses during personnel visits to OL1 and OL2 plants. Furthermore, we considered from a holistic perspective what such modelling would require for production use.

The study demonstrated that, in terms of the data used, doses measured by electronic dosimeters at the time of visits are suitable for analysis. Specifically, the radiation types measured in this context are γ (gamma) and X-ray (X-ray) radiation, which are collectively measured in microsieverts and classified as neutral radiations, on which the modelling was consequently based. These types of radiation are considered as indirect processes, meaning they do not directly ionise matter.

The study found that the resulting accumulation of radiation doses can be considered from a stochastic point of view, especially in terms of genetics. The assessment of radiation effects at low doses, however, poses significant challenges due to a lack of data. Partly because of these potential effects, the whole process, from the measurement of the doses themselves to the actual visit, is closely controlled and regulated by both TVO and STUK.

The data analysed in this study proved to be of a high quality, with the exception of a few weaknesses, but nevertheless not sufficiently informative to provide reliable predictions. Consequently, it was determined that the data requires further refinement from alternative data sources. To this end, five different machine learning models

were considered: 1) Random Forest, 2) Balanced Random Forest, 3) XGBoost, 4) LightGBM and 5) Easy Ensemble with AdaBoost. Of these, LightGBM and Random Forest emerged as the most effective models. However, each model had its limitations, with LightGBM demonstrating better overall performance and Random Forest indicating class-specific strengths.

The findings of this study indicate a need for further research, particularly in the area of data collection. It has been observed that enhancing the data might be feasible, though this can be quite challenging in certain instances. Additionally, it would be beneficial to explore a more comprehensive layout that incorporates data also from other than the annual maintenance periods. In such cases, the utilisation of deep learning systems could be a relevant approach. The use of real-time data for model maintenance is another essential aspect that, due to the study's limited time and scope, was not able to be explored. This highlights the complex nature and significant time demands of the subject.

However, the OL3 plant is in the process of generating similar data, and due to the plant's design, the accumulated doses will be lower, resulting in reduced variability. Theoretically, this should allow prediction at smaller class intervals without manipulating the data. In particular, it would be beneficial to develop a model that compares the OL3 data with that generated at OL1 and OL2, with the objective of achieving better results. However, this is only a future possibility, as at the time of writing, there is only data from the first year of operation.

This research highlights the key problem with AI solutions, which is data dependency. The emphasis is on quality over quantity, data must tell a story and not just repeat itself. In similar projects, the first step is always to identify the needs and the capabilities to meet those needs, because an AI solution can only be as good as the data applied to it.

References

- [1] O. A. Osibote, *Ionizing and Non-ionizing Radiation*. London, United Kingdom: IntechOpen, 2020, pp. 1–10, 101–116. DOI: 10.5772/intechopen.77474.
- [2] H. Domenech, *Radiation Safety: Management and Programs*. Miami, FL, USA: Springer International Publishing AG, 2016, pp. 1–71, 111–118, 143–182, 259–274. DOI: 10.1007/978-3-319-42671-6.
- [3] R. L. Murray and K. E. Holbert, *Nuclear Energy - An Introduction to the Concepts, Systems, and Applications of Nuclear Processes (7th Edition)*. San Diego, CA, USA: Elsevier, 2015, pp. 5–6, 72–82, 89–98, 139–149, 155–163. DOI: 10.1016/C2012-0-02697-X.
- [4] M. Hurlbert, L. Shasko, J. Condor, and D. Landrie-Parker, “Radiation Workers and Risk Perceptions: Low Dose Radiation, Nuclear Power Production, and Small Modular Nuclear Reactors”, *Journal of Nuclear Engineering*, vol. 4, pp. 258–277, 2023. DOI: 10.3390/jne4010020.
- [5] Teollisuuden Voima Oyj. “OL1 & OL2 Ydinvoimalaitosyksiköt”. (2023), [Online]. Available: https://tvo.fi/material/sites/tvo/pdf/k9su4vcbz/OL1_ja_OL2_laitosyksikot._Tekninen_esite.pdf (visited on 09/05/2024).
- [6] Radiation and Nuclear Safety Authority. “Radiation Protection and Exposure Monitoring of Nuclear Facility Workers”. YVL C.2. (2019), [Online]. Available: <https://www.stuklex.fi/en/ohje/YVLC-2> (visited on 08/12/2024).

-
- [7] “Assessment of Compliance with YVL Instructions C.2”, Teollisuuden Voima Oyj, Tech. Rep. 160021, 2020, OL1/OL2 (unpublished).
- [8] *General Principles for the Radiation Protection of Workers*. Oxford, United Kingdom: International Commission on Radiological Protection (ICRP), 1997, pp. 5–27, 43–46, 49–51. DOI: 10.1016/S0146-6453(97)88275-9.
- [9] S. Marsland, *Machine Learning: An Algorithmic Perspective, Second Edition*. Boca Raton, FL, USA: CRC Press LLC, 2014, pp. 6–11, 6–11, 22–25, 8–458. DOI: 10.1201/b17476.
- [10] Q. An, S. Rahman, J. Zhou, and J. J. Kang, “A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges”, *Sensors*, vol. 23, no. 9, 2023. DOI: 10.3390/s23094178.
- [11] M. R. Berthold, C. Borgelt, F. Höppner, and F. Klawonn, *Guide to Intelligent Data Analysis: How to Intelligently Make Sense of Real Data*. London, United Kingdom: Springer Nature, 2010, pp. 1–35, 97–105. DOI: 10.1007/978-1-84882-260-3.
- [12] Teollisuuden Voima Oyj. “Ydinvoimalaitosyksikkö Olkiluoto 3”. (2023), [Online]. Available: https://www.tvo.fi/material/sites/vanhattvo/20220825132746/7bmHsNHjV/ydinvoimalaitosyksikko_ol3_fin.pdf (visited on 09/05/2024).
- [13] Teollisuuden Voima Oyj. “Nuclear power plant units Olkiluoto 1 and Olkiluoto 2”. (2007), [Online]. Available: <https://www.tvo.fi/uploads/File/nuclear-power-plant-units.pdf> (visited on 09/13/2024).
- [14] “General Procedure for Radiation Protection”, *Radiation Protection Manual*, Teollisuuden Voima Oyj, Tech. Rep. 103296, 2019, OL1/OL2/OL3 (unpublished).

-
- [15] Ministry of Justice. “Radiation Act 9.11.2018/859”. (2018), [Online]. Available: <https://www.finlex.fi/en/laki/kaannokset/2018/en20180859> (visited on 08/12/2024).
- [16] Ministry of Justice. “Government Decree on Ionizing Radiation 22.11.2018/1034”. (2018), [Online]. Available: <https://finlex.fi/en/laki/kaannokset/2018/en20181034> (visited on 08/12/2024).
- [17] “Activities in the Controlled Area”, *Radiation Protection Manual*, Teollisuuden Voima Oyj, Tech. Rep. 138101, 2024, OL1/OL2/OL3/Posiva (unpublished).
- [18] “Operation and Quality Assurance of the TL Dosimeter System”, *Radiation Protection Manual*, Teollisuuden Voima Oyj, Tech. Rep. 105092, 2023, General (unpublished).
- [19] “Portable Radiation Monitoring Equipment Final Safety Datasheet”, Teollisuuden Voima Oyj, Tech. Rep. 107779, 2024, 556/JYC - OL1/OL2/OL3 (unpublished).
- [20] “User Instructions for Work Dosimeter System”, *Radiation Protection Manual*, Teollisuuden Voima Oyj, Tech. Rep. 139765, 2021, OL1/OL2/OL3 (unpublished).
- [21] Radiation and Nuclear Safety Authority. “Radiation and Nuclear Safety Authority Regulation on Measurements of Ionizing Radiation”. Annexes 1 and 2. (2021), [Online]. Available: <https://www.stuklex.fi/en/maarays/stuk-s-7-2021> (visited on 09/13/2024).
- [22] MPG Instruments. “DMC2000S Operating Manual”. (2021), [Online]. Available: https://ps-irrad.web.cern.ch/ps-irrad/assets/doc/info/DMC2000S_Operating_Manual.pdf (visited on 09/12/2024).

- [23] Mirion Technologies. “DMC3000 Data Sheet”. (2023), [Online]. Available: https://assets-mirion.mirion.com/prod-20220822/cms4_mirion/files/pdf/spec-sheets/dmc-3000-personal-electronic-dosimeter-data-sheet.pdf (visited on 09/12/2024).
- [24] “Mirion DMC 2000S SA Dosimeter Inspection and Operation Manual”, *Maintenance Manual*, Teollisuuden Voima Oyj, Tech. Rep. 139928, 2023, OL1/OL2/OL3 (unpublished).
- [25] “Mirion DMC-3000 Electronic Dosimeter Inspection and Operation Manual”, *Maintenance Manual*, Teollisuuden Voima Oyj, Tech. Rep. 179876, 2023, OL1/OL2/OL3 (unpublished).
- [26] “Alarm Limits in Electronic Dosimeters”, *Radiation Protection Manual*, Teollisuuden Voima Oyj, Tech. Rep. 119131, 2023, OL1/OL2/OL3/Posiva (unpublished).
- [27] “ALARA Program”, *Radiation Protection Manual*, Teollisuuden Voima Oyj, Tech. Rep. 108286, 2023, General (unpublished).
- [28] *Dose Control at Nuclear Power Plants: (Report No. 120)*. Bethesda, MD, USA: National Council on Radiation Protection and Measurements (NCRP), 1994, pp. 1–20, 46–51, 57–61, 88–95. [Online]. Available: <https://app.knovel.com/hotlink/toc/id:kpDCNPPRN2/dose-control-at-nuclear/dose-control-at-nuclear>.
- [29] W. R. Hendee and F. Marc Edwards, “ALARA and an Integrated Approach to Radiation Protection”, *Seminars in Nuclear Medicine*, vol. 16, pp. 142–150, 1986. DOI: [https://doi.org/10.1016/S0001-2998\(86\)80027-7](https://doi.org/10.1016/S0001-2998(86)80027-7).
- [30] International X-Ray and Radium Protection Committee or Commission (IXRPC), “International Recommendations for X-ray and Radium Protection”, *British*

- Journal of Radiology*, vol. 1, pp. 358–363, 2014. DOI: 10.1259/0007-1285-1-10-358.
- [31] “Use of the Radiation Work Permit”, *Radiation Protection Manual*, Teollisuuden Voima Oyj, Tech. Rep. 105107, 2023, General (unpublished).
- [32] J. Marttila, “Ydinenergian käytön turvallisuusvalvonta: Vuosiraportti 2023”, *STUK-B*, vol. 315, pp. 1–86, 2023. [Online]. Available: <https://urn.fi/URN:ISBN:978-952-309-597-7>.
- [33] V. Sinikka, V. Vesa-Pekka, T. Jani, T. Mikko, T. Tiina, and M. Aleks, “Monitoring of Radioactivity in the Environment of Finnish Nuclear Power Plants: Annual Report 2023”, *STUK-B*, vol. 318, pp. 1–47, 2023. [Online]. Available: <https://urn.fi/URN:ISBN:978-952-309-598-4>.
- [34] C. Schroerer, F. Kruse, and J. M. Gomez, “A Systematic Literature Review on Applying CRISP-DM Process Model”, *Procedia Computer Science*, vol. 181, pp. 526–534, 2021. DOI: 10.1016/j.procs.2021.01.199.
- [35] J. S. Saltz, “CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps”, in *2021 IEEE International Conference on Big Data (Big Data)*, New York, NY, USA: IEEE, 2021, pp. 2337–2344. DOI: 10.1109/BigData52589.2021.9671634.
- [36] S. Studer, T. B. Bui, C. Drescher, *et al.*, “Towards CRISP-ML(Q): A Machine Learning Process Model with Quality Assurance Methodology”, *Machine Learning and Knowledge Extraction*, vol. 3, pp. 392–413, 2021. DOI: 10.3390/make3020020.
- [37] Ministry of Justice. “Nuclear Energy Decree 12.2.1988/161”. (1988), [Online]. Available: <https://www.finlex.fi/en/laki/kaannokset/1988/en19880161> (visited on 08/12/2024).

- [38] Ministry of Justice. “Nuclear Energy Act 11.12.1987/990”. (1987), [Online]. Available: <https://www.finlex.fi/en/laki/kaannokset/1987/en19870990> (visited on 08/12/2024).
- [39] Ministry of Justice. “Ministry of Social Affairs and Health Decree on Ionising Radiation 22.11.2018/1044”. (2018), [Online]. Available: <https://www.finlex.fi/en/laki/kaannokset/2018/en20181044> (visited on 08/12/2024).
- [40] K. Janne, “Chernobyl Accident as a Turning Point of the Developing of the Radiation Monitoring System: Radiation Monitoring Before and After Chernobyl Accident”, PhD Thesis, University of Eastern Finland, 2021, pp. 28–30. [Online]. Available: <https://urn.fi/URN:ISBN:978-952-61-3736-0>.
- [41] “Personal Dosage Control and Implementation of Health Examination”, *Radiation Protection Manual*, Teollisuuden Voima Oyj, Tech. Rep. 104246, 2023, OL1/OL2/OL3/Posiva (unpublished).
- [42] “Work Dose Codes”, *Radiation Protection Manual*, Teollisuuden Voima Oyj, Tech. Rep. 112406, 2023, OL1/OL2/Posiva (unpublished).
- [43] International Commission on Radiological Protection (ICRP), “The 2007 Recommendations of the International Commission on Radiological Protection”, *Annals of the ICRP*, vol. 37, pp. 49–58, 2007. DOI: 10.1016/j.icrp.2007.10.003.
- [44] U.S. Nuclear Regulatory Commission. “High Radiation Doses”. (2020), [Online]. Available: <https://www.nrc.gov/about-nrc/radiation/health-effects/high-rad-doses.html> (visited on 10/07/2024).
- [45] Y. J. Kim, J. W. Lee, Y. H. Cho, Y. J. Choi, Y. Lee, and H. W. Chung, “Chromosome Damage in Relation to Recent Radiation Exposure and Radiation Quality in Nuclear Power Plant Workers”, *Toxics*, vol. 10, no. 94, 2022. DOI: 10.3390/toxics10020094.

- [46] M. Gillies, D. B. Richardson, E. Cardis, *et al.*, “Mortality from Circulatory Diseases and Other Non-cancer Outcomes Among Nuclear Workers in France, the United Kingdom and the United States (INWORKS)”, *Radiation Research*, vol. 188, pp. 276–290, 2017. DOI: 10.1667/rr14608.1.
- [47] T. Azizova, E. Grigoryeva, G. Zhuntova, E. Kirillova, and C. Loffredo, “Database of Families of Workers Chronically Exposed to Radiation: Data and Biospecimen Resources”, *Health Physics*, vol. 120, pp. 201–211, 2021. DOI: 10.1097/HP.0000000000001300.
- [48] A. Hasegawa, K. Tanigawa, A. Ohtsuru, *et al.*, “Health Effects of Radiation and Other Health Problems in the Aftermath of Nuclear Accidents, with an Emphasis on Fukushima”, *The Lancet*, vol. 386, pp. 479–488, 2015. DOI: 10.1016/S0140-6736(15)61106-0.
- [49] C. Song, T. Y. Kong, S. Kim, *et al.*, “High-Radiation-Exposure Work in Korean Pressurized Water Reactors”, *Nuclear Engineering and Technology*, vol. 5, pp. 1874–1879, 2024. DOI: 10.1016/j.net.2023.12.048.
- [50] V. V. Kosterev, A. G. Tsov’yanov, A. G. Sivenkov, and Y. N. Bragin, “Worker Radiation Exposure”, *Atomic Energy (New York, N.Y.)*, vol. 120, pp. 148–152, 2016. DOI: 10.1007/s10512-016-0110-2.
- [51] M. S. Pathan, S. M. Pradhan, T. P. Selvam, and B. K. Sapra, “A Multi-stage Machine Learning Algorithm for Estimating Personal Dose Equivalent Using Thermoluminescent Dosimeter”, *Machine Learning: Science and Technology*, vol. 5, no. 9, 2024. DOI: 10.1088/2632-2153/ad1c31.
- [52] M. Sasaki and Y. Sanada, “Improvement of Training Data for Dose Rate Distribution Using an Artificial Neural Network”, *Journal of Advanced Simulation in Science and Engineering*, vol. 9, pp. 30–39, 2022. DOI: 10.15748/jasse.9.30.

-
- [53] H. Breitzkreutz, J. Mayr, M. Bleher, S. Seifert, and U. Stöhlker, “Identification and Quantification of Anomalies in Environmental Gamma Dose Rate Time Series Using Artificial Intelligence”, *Journal of Environmental Radioactivity*, vol. 259–260, no. 9, 2023. DOI: 10.1016/j.jenvrad.2022.107082.
- [54] Y. Park, J. H. Choi, J.-B. Choi, and M. K. Kim, “A Stress Intensity Predictive Model for Reactor Pressure Vessel via Coupled Signal Processing and Machine Learning Model”, *Journal of Mechanical Science and Technology*, vol. 37, pp. 2881–2890, 2023. DOI: 10.1007/s12206-023-0514-6.
- [55] P. Ramirez-Hereza, D. Ramos, D. T. Toledano, J. Gonzalez-Rodriguez, A. Ariza-Velazquez, and N. Doncel, “Score-based Bayesian Network Structure Learning Algorithms for Modeling Radioisotope Levels in Nuclear Power Plant Reactors”, *Chemometrics and Intelligent Laboratory Systems*, vol. 237, no. 5, 2023. DOI: 10.1016/j.chemolab.2023.104811.
- [56] S. A. Balanya, D. Ramos, P. Ramirez-Hereza, *et al.*, “Gaussian Processes for Radiation Dose Prediction in Nuclear Power Plant Reactors”, *Chemometrics And Intelligent Laboratory Systems*, vol. 230, no. 19, 2022. DOI: 10.1016/j.chemolab.2022.104652.
- [57] M. K. Baek, Y. S. Chung, S. Lee, I. Kang, J. J. Ahn, and Y. H. Chung, “Design of a Nuclear Monitoring System Based on a Multi-sensor Network and Artificial Intelligence Algorithm”, *Sustainability*, vol. 15, no. 7, 2023. DOI: 10.3390/su15075915.
- [58] S. Theodoridis and K. Koutroumbas, *Pattern Recognition, Fourth Edition*. Burlington, MA, USA: Academic Press, 2009, pp. 1–9, 262–265, 323–326, 570–577. DOI: 10.1016/B978-1-59749-272-0.X0001-2.
- [59] S. Raschka, “Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning”, *ArXiv*, pp. 1–49, 2018. DOI: 10.48550/arXiv.1811.12808.

-
- [60] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System”, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: Association for Computing Machinery, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785.
- [61] G. Ke, Q. Meng, T. Finley, *et al.*, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree”, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, Red Hook, NY, USA: Curran Associates, Inc., 2017, pp. 3149–3157. [Online]. Available: <http://papers.nips.cc/paper/6907-lightgbm-a-highly-efficient-gradient-boosting-decision-tree.pdf>.
- [62] L. Breiman, “Random Forests”, *Machine Learning*, vol. 45, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324.
- [63] C. Chen, A. Liaw, and L. Breiman, “Using Random Forest to Learn Imbalanced Data”, pp. 1–12, 2004. [Online]. Available: <https://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf>.
- [64] X.-Y. Liu, J. Wu, and Z.-H. Zhou, “Exploratory Undersampling for Class-Imbalance Learning”, *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, pp. 539–550, 2009. DOI: 10.1109/TSMCB.2008.2007853.
- [65] P. Virtanen, R. Gommers, T. E. Oliphant, *et al.*, “SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python”, *Nature Methods*, vol. 17, pp. 261–272, 2020. DOI: 10.1038/s41592-019-0686-2.
- [66] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, “Scikit-learn: Machine Learning in Python”, *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. DOI: 10.48550/arXiv.1201.0490.

- [67] C. R. Harris, K. J. Millman, S. J. van der Walt, *et al.*, “Array Programming with NumPy”, *Nature*, vol. 585, pp. 357–362, 2020. DOI: 10.1038/s41586-020-2649-2.
- [68] W. McKinney, “Data Structures for Statistical Computing in Python”, in *Proceedings of the 9th Python in Science Conference*, Austin, TX, USA: SciPy, 2010, pp. 56–61. DOI: 10.25080/Majora-92bf1922-00a.
- [69] J. D. Hunter, “Matplotlib: A 2D Graphics Environment”, *Computing in Science & Engineering*, vol. 9, pp. 90–95, 2007. DOI: 10.1109/MCSE.2007.55.
- [70] S. Galli, *Python Feature Engineering Cookbook, Second Edition*. Birmingham, United Kingdom: Packt Publishing, Limited, 2022, pp. 1–358, ISBN: 978-1804611302.
- [71] G. Bontempi, S. Ben Taieb, and Y.-A. Le Borgne, “Machine Learning Strategies for Time Series Forecasting”, in *Business Intelligence: Second European Summer School, eBISS 2012, Brussels, Belgium, July 15-21, 2012, Tutorial Lectures*. Berlin, Germany: Springer Berlin Heidelberg, 2013, pp. 62–77. DOI: 10.1007/978-3-642-36318-4_3.
- [72] M. P. Ptotic, M. B. Stojanovic, and P. M. Popovic, “A Review of Machine Learning Methods for Long-Term Time Series Prediction”, in *2022 57th International Scientific Conference On Information, Communication And Energy Systems And Technologies (ICEST)*, Ohrid, North Macedonia: IEEE, 2022, pp. 205–208. DOI: 10.1109/ICEST55168.2022.9828618.
- [73] H. Kaur, H. S. Pannu, and A. K. Malhi, “A Systematic Review on Imbalanced Data Challenges in Machine Learning: Applications and Solutions”, *ACM Computing Surveys*, vol. 52, pp. 1–36, 2019. DOI: 10.1145/3343440.
- [74] A. Gossmann. “Probabilistic interpretation of AUC”. (2018), [Online]. Available: <https://www.alexejgossmann.com/auc/> (visited on 10/08/2024).

-
- [75] T. Fawcett, “An introduction to ROC analysis”, *Pattern Recognition Letters*, vol. 27, pp. 861–874, 2006. DOI: doi.org/10.1016/j.patrec.2005.10.010.
- [76] A. Paleyes, R.-G. Urma, and N. D. Lawrence, “Challenges in Deploying Machine Learning: A Survey of Case Studies”, *ACM Computing Surveys*, vol. 55, pp. 1–29, 2023. DOI: [10.1145/3533378](https://doi.org/10.1145/3533378).
- [77] J. Chandrasekaran, C. Tyler, N. McCarthy, E. Lanus, and L. Freeman, “Test & Evaluation Best Practices for Machine Learning-enabled Systems”, *arXiv.org*, pp. 1–20, 2023. DOI: [10.48550/arXiv.2310.06800](https://doi.org/10.48550/arXiv.2310.06800).
- [78] N. Polyzotis, S. Roy, S. E. Whang, and M. Zinkevich, “Data Lifecycle Challenges in Production Machine Learning: A Survey”, *Sigmoid Record*, vol. 47, pp. 17–28, 2018. DOI: [10.1145/3299887.3299891](https://doi.org/10.1145/3299887.3299891).
- [79] R. Mukhamediev I, Y. Popova, Y. Kuchin, *et al.*, “Review of Artificial Intelligence and Machine Learning Technologies: Classification, Restrictions, Opportunities and Challenges”, *Mathematics*, vol. 10, no. 15, 2022. DOI: [10.3390/math10152552](https://doi.org/10.3390/math10152552).
- [80] U. Bhatt, A. Xiang, S. Sharma, *et al.*, “Explainable Machine Learning in Deployment”, in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain: Association for Computing Machinery, 2020, pp. 648–657. DOI: [10.1145/3351095.3375624](https://doi.org/10.1145/3351095.3375624).