République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université Batna 2 – Mostefa Ben Boulaïd
Faculté de Technologie
Département d'Électronique

Thèse

Préparée au sein du Laboratoire des Systèmes Propulsion - Induction Électromagnétiques

Présentée pour l'obtention du diplôme de :

Doctorat 3ème Cycle LMD en Électronique
Option : Électronique Médicale

Sous le Thème :

Multi-criteria optimization approaches for body radiation therapy

Présentée par :

MEDDOUR Abderrahim

Devant le jury composé de :

KOUDA Souhil	MCA.	Université de Batna 2	Président
DRID Said	Prof.	Université de Batna 2	Rapporteur
DENDOUGA Abdelghani	MCA.	Université de Batna 2	Examinateur
ABDI Mohamed Amir	MCA.	ENSEREDD*	Examinateur
BENDIB Toufik	MCA.	ENSEREDD*	Examinateur

2020/2021

* Ecole Nationale Supérieure En Energies Renouvelables, Environnement et Développement Durable

بِسْمِ ٱللَّهِ ٱلرَّحْمَٰنِ ٱلرَّحِيمِ

Acknowledgements

First of all, I thank ALLAH almighty who armed me with determination, patience and courage during all these years of study.

I would like to express my sincere gratitude to MCA. KOUDA Souhil for having agreed to preside this jury.

It is anything but easy to thank all the people who contributed to my personal and professional development during my Ph.D. studies. More than anyone else, my advisor, Pr. Said DRID, for his guidance and personal support throughout my doctoral study, for the privileges he gave me to work on various challenging and interesting areas. I am always impressed and inspired by his sharp insight, deep wisdom and profound knowledge.

I am also grateful to MCA. Mohamed Amir ABDI, MCA. Toufik BENDIB and MCA. Abdelghani DENDOUGA, who accepted to be my jury members, and devoted their precious time to review my thesis. I also want to thank my brother, MCA. Fayçal MEDDOUR, for his solid technical guidance, continuous support and encouragement in my work and for his responsiveness to my requests.

I would like to thank Dr. Toufik BENTERCIA for his orientations and his judicious advice.

A particular gratitude to Dr. Hichem BENCHERIF for his solid technical guidance, orientations and his judicious advice.

Dédication

To my father

For his patience and these considerable sacrifices to get me to this level.

To my mother

For her great love, these sacrifices and all the affectation she has always offered me.

To my brothers, my sisters

To my wife, my little daughter

To all my colleagues and friends.

I dedicate this Thesis.

Abstract

Radiation therapy or Radiotherapy is one of the therapeutic uses of ionizing radiation to treat cancer. Depending on the type of tumor and its location, different modes of radiotherapy are used in the clinic, Brachytherapy, Metabolic radiotherapy and External radiotherapy which is the most common form of radiotherapy. This latter consists of irradiating tumor location from an external radiation source such medical linear accelerators LINACs. Many external radiotherapy techniques have been developed, the most important called intensity modulated radiation therapy IMRT.

In radiotherapy, the goal has always focused on achieving a conformal radiation therapy plan, where a high dose of radiation conforms to the tumor, while the radiation unavoidably received to the surrounding healthy organs and tissues is minimized. The investigation of such goal is the main topic of this PhD document. In particular, the dissertation concerns the optimization of intensity modulated radiation therapy (IMRT) treatment planning by solving the fluence map optimization (FMO) problem using multiobjective genetic algorithm MGA. This, offers to the oncologists and physicians a set of conformal treatment plans.

Moreover, to ensure a safe patient dose verification, a junctionless double graphene gate radiation sensitive FET (RADFET) is proposed as dosimeter. besides associated analytical analysis are both introduced. In addition, the effect of graphene work function on the device performance measures is also investigated. Furthermore, the elaborated model defines the figures of merit in the context of (MGA) technique. The improved electrical response is compared with existing double gate (DG) RADFETs, where the proposed device figures of merit reveal that the optimized proposed RADFET provides improved electrical performance and sensitivity, and therefore, enhancing radiation therapy quality assurance QA.

Résumé

La radiothérapie est l'une des utilisations thérapeutiques des rayonnements ionisants pour traiter le cancer. Selon le type de tumeur et sa localisation, différents modes de radiothérapie sont utilisés en clinique, la curiethérapie, la radiothérapie métabolique et la radiothérapie externe qui est la forme la plus courante de radiothérapie. Cette dernière consiste à irradier la localisation tumorale à partir d'une source de rayonnement externe tels les accélérateurs linéaires médicaux LINACs. De nombreuses techniques de radiothérapie externe ont été développées, la plus importante étant appelée radiothérapie conformationnelle avec modulation d'intensité (RCMI). En radiothérapie, l'objectif s'est toujours concentré sur la réalisation d'un plan de radiothérapie conforme, où une dose élevée de rayonnement se conforme à la tumeur, tandis que le rayonnement inévitablement reçu vers les organes et tissus sains environnants est minimisé. L'investigation d'un tel objectif est le sujet principal de ce document de thèse. En particulier, la thèse concerne l'optimisation de la planification du traitement par radiothérapie conformationnelle avec modulation d'intensité en résolvant le problème d'optimisation de la carte de fluence (FMO) à l'aide de l'algorithme génétique multiobjectif (MGA). Celui-ci offre aux médecins oncologues et aux physiciens un ensemble de plans de traitement conformes. De plus, pour assurer une vérification sûre de la dose au patient, un FET sensible au rayonnement à double grille de graphène sans jonction (RADFET) est proposé comme dosimètre. En plus une analyse analytique associée sont tous deux introduits. En outre, l'effet de la fonction de travail du graphène sur les mesures de performance du dispositif est également étudié. De plus, le modèle élaboré définit les figures de mérite dans le cadre de la technique (MGA). La réponse électrique améliorée est comparée aux RADFET à double porte (DG) existants, où les chiffres de mérite du dispositif proposé révèlent que le RADFET proposé optimisé offre des performances et une sensibilité électriques améliorées, et donc, améliore l'assurance de qualité de la radiothérapie.

ملخص

العلاج بالأشعة أو العلاج الإشعاعي هو أحد الاستخدامات العلاجية للأشعة المؤينة لعلاج السرطان. اعتمادًا على نوع الورم وموقعه، يتم استخدام طرق مختلفة من العلاج الإشعاعي في العيادة: العلاج الإشعاعي الموضعي، العلاج الإشعاعي الأيضي والعلاج الإشعاعي الخارجي وهو أكثر أشكال العلاج الإشعاعي شيوعًا. يعتمد هذا الأخير على تعريض موقع الورم لأشعة من مصدر إشعاع خارجي مثل المسرعات الخطية الطبية LINACs. تم تطوير العديد من تقنيات العلاج الإشعاعي الخارجي وأهمها العلاج الإشعاعي المعدل الشدة IMRT.

في العلاج الإشعاعي، ركز الهدف دائمًا على تحقيق خطة علاج إشعاعي ملائمة، حيث تتوافق جرعة عالية من الإشعاع مع الورم، بينما يتم تقليل الإشعاع الذي يتم تلقيه بشكل لا مفر منه للأعضاء والأنسجة السليمة المحيطة. تحقيق هذا الهدف هو الموضوع الرئيسي لأطروحة الدكتوراه هذه. على وجه الخصوص، تتعلق الرسالة بتحسين مخطط العلاج الإشعاعي المعدل الكثافة (IMRT) من خلال حل مشكلة تحسين خرائط الطلاقة (FMO) باستخدام الخوارزمية الجينية متعددة الأغراض MGA. وهذا يقدم لأطباء الأورام مجموعة من خطط العلاج الملائمة.

علاوة على ذلك، لضمان التحقق الآمن من جرعة المريض، اقترحنا استخدام ترانزستور بتأثير الحقل (FET) ببوابة الجرافين المزدوجة الحساسة للإشعاع كمقياس للجرعات. تم تقديمه إلى جانب التحليل التحليلي المرتبط به. بالإضافة إلى ذلك، تم أيضًا فحص تأثير وظيفة عمل الجرافين على مقاييس أداء الجهاز. إضافة إلى ذلك ، يحدد النموذج المفصل أرقام الجدارة في سياق تقنية (MGA). تتم مقارنة الاستجابة الكهربائية المحسّنة مع RADFETs ذات البوابة المزدوجة (DG) الحالية، حيث تكشف أرقام الجدارة المقترحة للجهاز أن RADFET المُحسَّن المقترح يوفر أداء وحساسية كهربائية مُحسَّنتيين، وبالتالي يعزز ضمان جودة العلاج الإشعاعي QA.

Contents

Abbreviations and common symbols

Abbreviation	Full definition	Introduced in
3D-CRT	Three-Dimensional Conformal Radiation Therapy	I.6.2
APC	Anomalous Positive Charge	II.3.5
BAO	Beam Angle Optimization	I.11
BEV	Beam Eye View	I.6.2
CERR	computational environment for radiotherapy research	I.8.1
CP	Charge-Pumping	II.4.2
CT	Computerized tomography	I.6.2
CTV	Clinical Target Volume	I.5.1
DAO	Direct Aperture Optimization	I.11
DMLC	Dynamic MultiLeaf Collimation	I.6.3.2
DNA	Double Strand Breaks	I.2
DVH	Dose-Volume Histogram	I.5.2
EAs	Evolutionary Algorithms	III.1
EMO	Evolutionary Multiobjective Optimization	III.1
ENT	Ear Nose Throat	I.6.2
FMO	Fluence Map Optimization	I.11
FST	fast switching traps	II.3.5
FT	Fixed Traps	II.3.5
GA	Genetic Algorithm	III.4.3
GAA MOSFET	Gate All Around Metal Oxide Semiconductor FET	V.1
GOL	Gate Oxide Layer	II.2.4
GTV	Gross Tumor Volume	I.5.1
ICRU	International Commission on. Radiation Units	I.5.1
IMRT	Intensity Modulated Radiation Therapy	I.6.3
JL DG FET	Junction-less Double Gate Field Effect Transistor	V.1
LINAC	Linear Accelerator	I.4
LP	Linear Programming	I.11.3.5
MC	Monte Carlo algorithms	I.9.1
MEMS	Microelectromecanicanical Systems	II.2.3
MG	Midgap-subthreshold	II.4.1
MGA	Multiobjective Genetic Algorithm	III.4.3
MLC	MultiLeaf Collimator	I.6.1
MLCSO	MLC Segmentation Optimization	I.11
MO	Multiobjective Optimization	IV.1
MOP	Multiobjective Problem	III.1
MOS	Metal Oxide Semiconductor	II.2.2
MRI	Magnetic Resonance Imaging	I.5.1
MIP	Mixed-Integer Programming	I.11.3.4
NBO	Non-Bridging Oxygen	II.3.2
NLP	Non-Linear Programming	I.11.3.3
NSGA	Non-Sorted Genetic Algorithm	III.4.3
OAR	Organ At Risk	I.5.2
OSLDs	Optically stimulated luminescence dosimeters	II.2.1

PAES	Pareto archived evolution strategy	III.4.4
PET	Positron Emission Tomography	I.7
PTV	Planning Target Volume	I.5.1
PRV	Planning Risk Volume	I.5.2
QA	Quality Assurance	I.8
QIB	Quadrant Infinite Beam	I.9.4
RADFET	Radiation-Sensing Field-Effect-Transistors	II.2.4
RT	Radiation Therapy	I.1
SCH1	Schaffer's Problem No.1	III.4.4
SE	Secondary Electron	II.3.1
SMLC	Segmential MultiLeaf Collimation	I.6.3.1
SOT	Switching Oxide Traps	II.3.5
SPECT	Single Photon Emission Tomography	I.7
SS	slow states	II.3.5
SST	slow switching traps	II.3.5
ST	Switching Traps	II.3.5
TLD	Thermo-Luminescent Dosimeters	II.2.4
TPS	Treatment Planning System	I.7
VMAT	Volumetric Modulated Arc Therapy	I.6.4
WHO	World Health Organization	General introduction

General Introduction

Cancer ranks as a leading cause of death and an important barrier to increasing life expectancy in every country of the world [1]. According to estimates from the World Health Organization (WHO) in 2019 [2], cancer is the first or second leading cause of death before the age of 70 years in 112 of 183 countries. With cancer affecting more individuals on a yearly basis, Radiation Therapy (RT) has become a key component in the successful management of this illness, both with curative and palliative intent, with over fifty percent of patients receiving some form of RT. Based on the location of the tumor, severity of the disease, and health of the patient, RT can be used as the stand-alone treatment or in conjunction with other treatment modalities such as surgery and chemotherapy.

X-rays were used for cancer treatments long before the radiobiological effects of ionizing radiation on human cells were understood. The first therapeutic application of x-rays occurred in 1896, less than a year after their discovery by Wilhelm Roentgen in November of 1895 [3]. The discovery of radioactivity and radium in 1896 [4] and 1898 [5], respectively, led to additional treatment options for cancer patients. However, it was not until the work of Regaud [6] and Coutard [7] in the early 20th century which suggested that normal tissue cells may be better able to repair damage due to radiation than cancer cells. Another significant conclusion from their work was the discovery that delivering radiation through a course of smaller fractions, as opposed to the then-current practice of delivering the entire amount of radiation in one sitting, allows for the exploitation of the enhanced repair mechanism in healthy tissue. This led to a reduction in complications arising from RT-based treatments and the ability to escalate the dose to the tumor with an improved probability of eradicating the cancer while not incurring additional complications to surrounding healthy tissues. Even without this crucial knowledge, from the first treatments using RT, the goal has always focused on achieving a conformal RT plan, where a high dose of radiation conforms to the tumor, while the radiation unavoidably received to the surrounding healthy organs and tissues is minimized. Depending on the type of tumor and its location, different modes of radiotherapy are used in the clinic, Brachytherapy [8], Metabolic radiotherapy and External radiotherapy which is the most common form of radiotherapy. During the therapy, the patient sits or lies on a couch and an external source of radiation is pointed at a particular part of the body. Advances in radiation physics and computer technology with the appearance of a device known as

the MultiLeaf Collimator (MLC) [9] during the last quarter of the 20th century, made it possible to aim radiation more precisely through the development of many techniques. In the mid-1990s a technique known as intensity modulated radiation therapy (IMRT) emerged which further enables tailoring of the 3Ddose distribution inside the patient. All of this led to the emergence of what we call it Treatment Planning System (TPS). The latter appeared with the development of imaging and calculation codes in radiotherapy. This computer tool is particularly useful in the context of IMRT to advantageously use the many possibilities of irradiation protocols. The TPS helps throughout the processing chain: data acquisition, delineation of structures, definition of beams angles and their weighs, dose calculation and check of calculation / measurement concordance. In the context of this thesis, we were only interested in two elements of this chain: the definition of beams weights and check of calculation / measurement concordance.

After selecting suitable beams orientations, the next logical phase is determining their weights. In the IMRT optimization process, due to discretization each beam into small beamlets, this step becomes determining the optimal fluence pattern for each beam that will result in the best-possible dose distribution in relation to some predetermined prescription dose and dose constraints to the relevant structures. Determining these optimal fluence patterns for a fixed set of beams forms the basis of the fluence map optimization (FMO) that is investigated in this dissertation.

From Another side, it is important to be able to measure the radiation dose accurately and make sure the proper amount of radiation is delivered. The advancement of modern radiation therapy technology, such as IMRT make dose verification a more and more complicated problem. Here came the important role of Dosimeters. These latter have many different ways to use, including: Phantom measurements, Skin surface dosimeters, in vivo dosimeters…etc. Clearly, patient dose verification at the point of delivery is an important part of quality assurance in radiotherapy treatment [9]. When included as part of a general system of radiotherapy quality assurance within a clinic, Dosimeters can significantly reduce the risk of mistreatment. Many existing radiation dosimetry technologies, including Optically stimulated luminescence dosimeters (OSLDs), Metal Oxide Semiconductor (MOS) capacitor, Microelectromechanical (MEMS) technology and Radiation-Sensing Field-Effect-Transistor (RADFET) which is our concern due to its reliability and accuracy. The basic RADFET dosimeter principle relies on the calculation of the threshold voltage change followed by the conversion of such difference to the absorbed dosage. Due to their advantages over conventional dosimetry systems, different MOSFET based dosimeters have been

produced in recent decades, and several relevant contributions have been reported to boost the sensitivity of RADFET through considering gate stack pMOS characterized by two layers of gate oxide Dual Dielectric materials. It is worthy to mention that pMOS Dosimeter, GAA MOSFET and JL DG FET are actually recommended for radiation sensor owing to their processing benefits and high immunity to short-channel effects. In addition to sensitivity enhancement challenges, increasing electrical performance and extracting optimal RADFET model parameters is ranked also as an optimization problem. Both IMRT FMO and RADFET performance optimization problems lead us to the Artificial Intelligence (AI) exploration and exploitation of its computational techniques.

Among AI models, such as cellular automata, artificial neural networks, fuzzy systems, multiagent systems, and swarm intelligence, genetic algorithms (GAs; Holland, 1975) have proved to be an effective and robust support tool for the prediction and modeling of complex phenomena. GAs belong to the broader family of evolutionary algorithms (EAs) and can be considered as both artificial models of natural evolution and general-purpose search algorithms. In particular, in this latter form, GAs have been employed for optimizing a broad variety of problems for which standard optimization techniques require excessive computational resources and time to return the result or, simply, for those problems for which specific optimization procedures do not exist.

In our dissertation, Multiobjective Genetic Algorithm (MGA) is the suitable tool proposed to solve the IMRT FMO problem, due to conflicting treatment goals - delivering a maximum dose to the tumor while providing a minimum dose to the healthy structures. When dealing with multiobjective optimization problems, the concept of optimality is generally extended according to the notion Pareto optimality, and refers to finding good tradeoff solutions among all the objectives, because the latter are commonly in conflict with each other. In fact, multiobjective optimization problems generally do not have one single optimal solution (global optimum) but a set of feasible solutions, each one better with respect to one particular objective and not as good with respect to others. In a multiobjective optimization problem, a set of (non-dominated Pareto optimal) solutions is, thus, found instead of one single solution.

For the enhancing radiation therapy quality QA and ensure a safe patient dose verification, a junctionless double graphene gate radiation sensitive FET (RADFET) besides associated analytical analysis are both introduced. In addition, the effect of graphene work function on the device performance measures is also investigated. Moreover, the elaborated model defines the figures of

merit in the context of (MGA) technique. The improved electrical response is compared with existing double gate (DG) RADFETs, where the proposed device figures of merit reveal that the optimized proposed RADFET provides improved electrical performance and sensitivity.

Outline of thesis

This dissertation is made up of 5 chapters:

The first chapter, presents the state of art of radiation therapy including biological reaction, delivery modes, External RT principal, different techniques with focus on IMRT treatment planning modality, dose calculation methods and the IMRT optimization problem.

The second chapter presents the state of art of RADFET dosimeter and its crucial role in RT quality assurance, in addition to a short presentation for other existing radiation dosimetry technology.

The third chapter introduces the Artificial Intelligence techniques, and focus on MGA in order to apply it next for the optimization of both IMRT fluence map and junctionless double graphene gate radiation sensitive FET (RADFET). In this sense, all the theoretical elements necessary for such an optimization technique are clearly developed.

The fourth chapter is devoted to MGA optimization process for IMRT fluence map, a result dose distribution for a liver case is presented and discussed.

The last chapter is devoted to the junctionless double graphene gate radiation sensitive FET (RADFET) besides associated analytical analysis. Analytical models using the technique of variables separation are implemented to measure and evaluate the capabilities of both proposed and standard RADFET devices. In addition, the effect of graphene work function on the device performance measures is also investigated. Moreover, the elaborated model defines the figures of merit in the context of a multi-objective genetic algorithm (MGA) technique. The improved electrical response is compared with existing double gate (DG) RADFETs, where the proposed device figures of merit reveal that the optimized proposed RADFET provides improved electrical performance and sensitivity.

References

[1] Bray F, Laversanne M, Weiderpass E, Soerjomataram I. The ever-increasing importance of cancer as a leading cause of premature death worldwide. Cancer. In press.

[2] World Health Organization (WHO). Global Health Estimates 2020: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2019. WHO; 2020.

[3] J.M. Slater, "Ion Beam Therapy," 320, 3–17 (2012).

[4] A.H. Bécquerel, "Sur les radiations invisibles emises par les corps phosphorescents," C. R. Acad. Sci. Paris 122, 501 (1896).

[5] P. Curie, M. Curie, and G. Bémont, "Radium, A new body, strongly radio-active, contained in pitchblende," Sci. Am (1899).

[6] C. Regaud and R. Ferroux, "Discordance des effets des rayons X, d'une part dans la peau, d'autre part dans le testicule par le fractionement de la dose: diminution de l'efficacite dans," CR Soc. Biol (1927).

[7] H. Coutard, "Principles of x ray therapy of malignant diseases," Lancet 224(5784), 1–8 (1934).

[8] Chargari, C., Deutsch, E., Blanchard, P., Gouy, S., Martelli, H., Guérin, F., Brachytherapy: An overview for clinicians. CA: A Cancer Journal for Clinicians, Haie-Meder, C. (2019).

[9] Taskin, Z., Smith, J., Romeijn, H., & Dempsey, J. (2010). Optimal Multileaf Collimator Leaf Sequencing in IMRT Treatment Planning. Oper. Res., 58, 674-690.

Chapter I:

State of Art of Radiation Therapy

Chapter I: State of Art of Radiation Therapy

I.1. Introduction

Radiation therapy (RT) is one of the therapeutic uses of ionizing radiation. Its origins date back to the beginning of the century, after the discovery of X-rays by W. Röntgen (1895), of radioactivity by H. Becquerel (1896) and of radium 226. by P. and M. Curie (1898). More than half of newly diagnosed cancer cases are treated with this technique and almost 50% of cures are partly or totally due to radiotherapy. Radiotherapy is mainly used in oncology, to treat, in combination or not with surgery and / or chemotherapy, the primary tumor, satellite lymphadenopathy and often certain metastases (especially bone and brain). Modern radiotherapy developed from 1950 with the advent of high energy devices (telecobalts, linear accelerators) and the replacement of radium 226 by artificial radioelements (iridium 192 and cesium 137). Radiation therapy for cancer often has side effects. Some of these effects are unavoidable and often go away on their own or with treatment. Side effects may occur due to the reaction of sensitive normal tissues located near the treated area or, more rarely, due to a particularly high individual sensitivity to ionizing radiation.

The state of the art detailed in this section presents only the points essential to a good understanding of the work carried out. Numerous works provide more exhaustive information on dosimetry and treatment planning techniques, it is possible, for example, to refer to the works of Anders Ahnesjo et al [1].

The purpose of radiation therapy is to damage tumor cells in order to prevent them from reproducing or to destroy them. The objective is to determine what is the best treatment regimen to obtain the desired effects, i.e., the destruction of tumor cells while minimizing the side effects associated with the treatment (protection of healthy cells peripheral to the tumor).

I.2. Radiobiology

Ionizations and absorbed energy damage the double strand breaks (DNA) of cells in normal tissues and malignancies, preventing them from dividing and growing [2]. Although the radiation is focused towards the tumor, it is unavoidably absorbed by the surrounding normal tissues, causing harm. Normal tissues and malignancies, on the other hand, are vulnerable to distinct biological reactions to radiation. Radiation therapy is based on the differential action of ionizing radiations, which destroy tumor cells while maintaining healthy tissue to some extent due to normal cells' ability to repair DNA double strand breaks.

As mentioned before, in addition to the possibility of combining radiotherapy with other cancer treatment strategies such as; chemotherapy, surgery or immunotherapy, recently, there has been increasing interest in combining radiotherapy methods with drug compounds, in attempts to improve the chances of successfully targeting and killing cancer cells (Fig I.1).

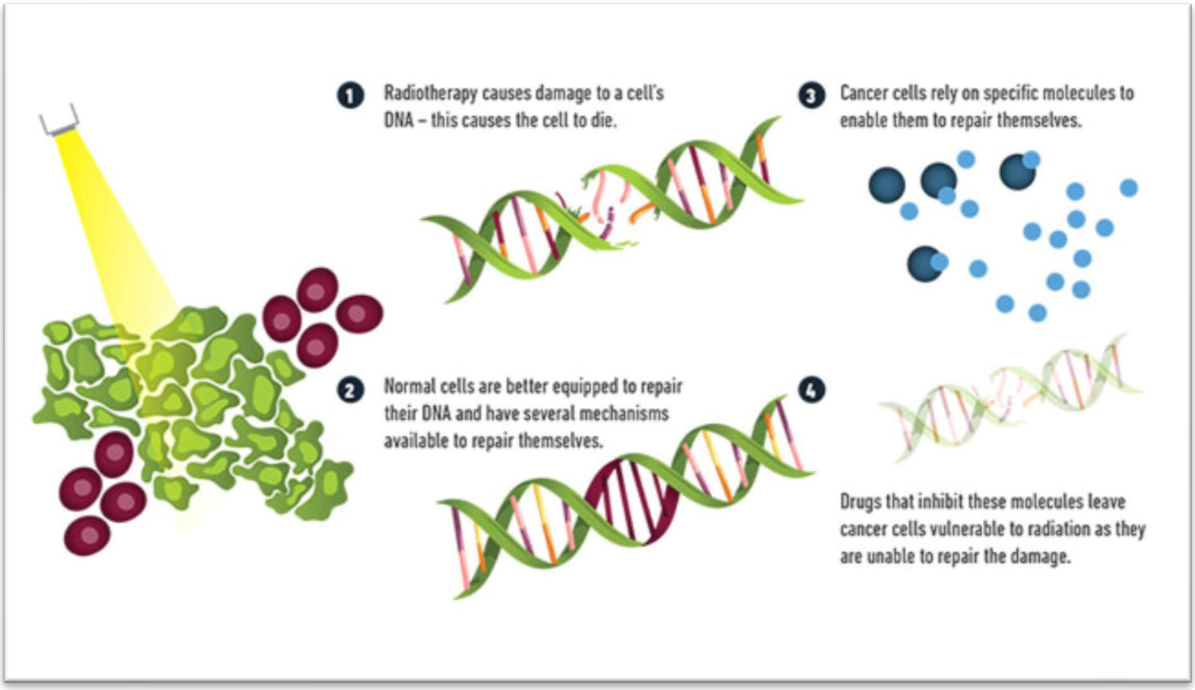


Fig.I.1. RT treatment mechanism and effects enhancement via drugs that prevent mechanisms of cancer cell repair.

I.3. RT delivery modes

Depending on the type of tumor and its location, different modes of radiotherapy are used in the clinic, shown in Figure I.2.

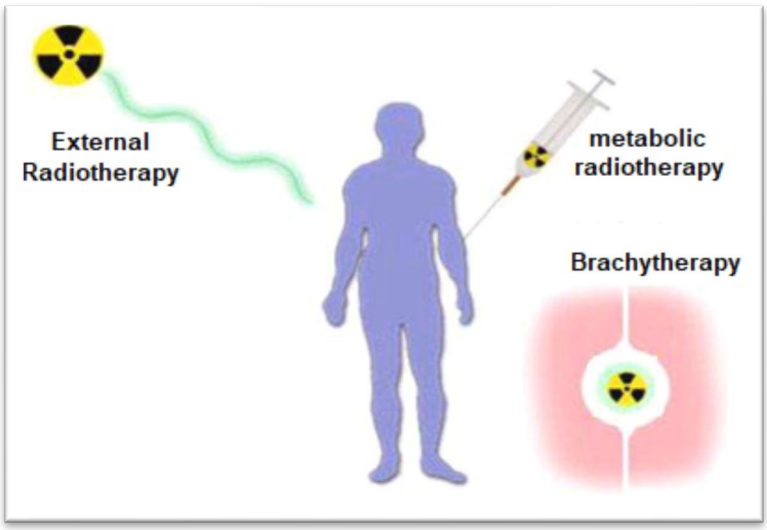


Fig.I.2. Three main modes of radiotherapy.

The application of irradiation, as a rule, can be done in several ways such as:

- External radiotherapy or Transcutaneous radiotherapy or Teleradiotherapy which uses beams of radiation penetrating the tissues through the skin. It is this method that is used in the context of this study.
- Brachytherapy involves introducing radioactive substances into the body by placing them in a natural hollow space, in the tumor itself or in its immediate vicinity. Irradiation is generally isotropic. It is therefore necessary to study the distribution of sources so that the tumor is properly destroyed while minimizing the impact on healthy tissues, which is the very principle of optimization in brachytherapy.
- Metabolic radiotherapy, which uses radioelements administered in liquid form. Metabolic radiation therapy is mostly used in some forms of thyroid cancer. In this case, the radioactive substance is administered orally or intravenously and will preferentially bind to cancer cells.

I.4. External radiotherapy

External beam radiation therapy is the most common. The source of radiation is outside the patient. It consists of administering the rays through the skin and tissues to irradiate the entire region affected by the tumor as well as possibly the nearest lymph nodes.

In external beam radiation therapy, irradiations are produced by a linear accelerator (see Figures I.3 and I.4). Radiation, depending on its type, can be directly (electrons) or indirectly (photons) ionizing. In both cases, the principle targeted is the destruction of the cancer cell.

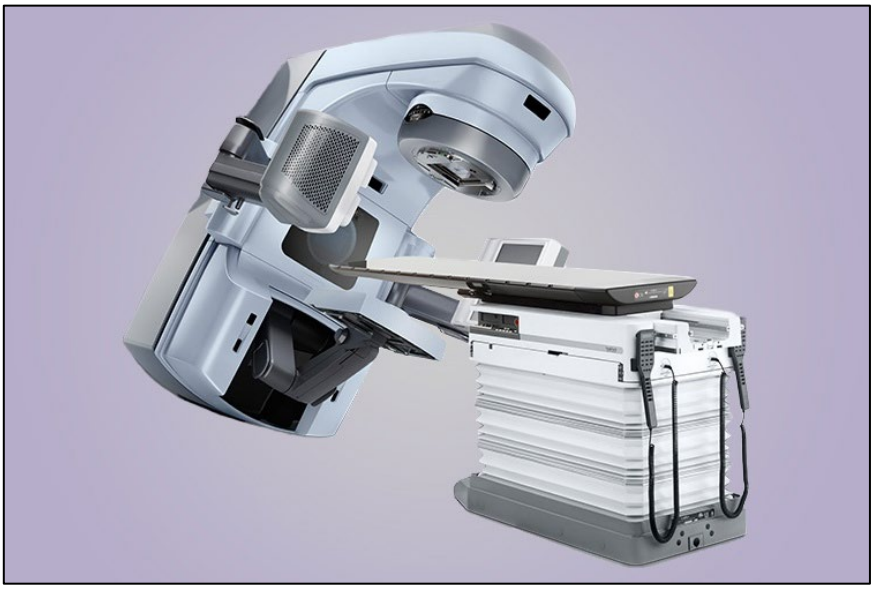


Fig.I.3. Medical Linear accelerator (Linac).

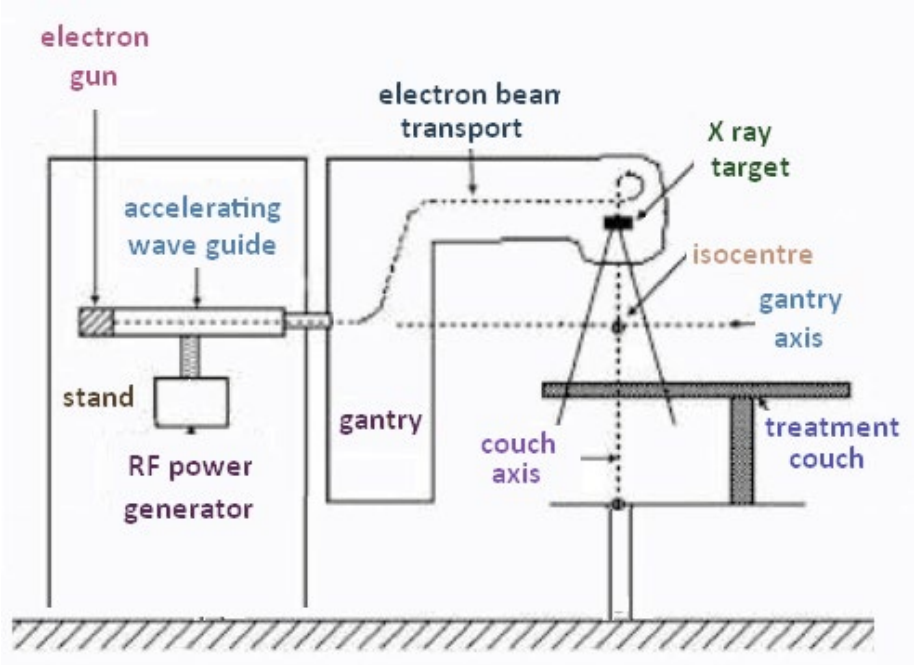


Fig.I.4. Medical linear accelerator principle.

The effect of treatment with ionizing radiation is measured as a function of the dose absorbed by the treated medium. The absorbed dose corresponds to the average energy deposited by the ionizing particles per unit mass of a material:

$$D = \frac{\overline{dEab}}{dm} \qquad \text{Unit: Gray } (Gy = J/kg) \qquad \text{(I.1)}$$

The absorbed dose is expressed in gray (Gy). One gray is equal to one Joule (J) of energy absorbed in one kilogram (kg) of matter. 2 Gy represents a daily dose of radiation which is generally tolerated by healthy cells. This feature is exploited for external radiotherapy by fractionation. For example, a total dose of 60 Gy can be delivered in fractions of 2 Gy over 30 days of treatment. For each treatment, the prescribed dose and its fractionation therefore depend on the location and nature of the disease.

The practical benefit of using the absorbed dose as a measurement unit for evaluating a treatment is that it is a purely physical measurement and therefore can be verified using a dosimeter.

In the context of a photon arriving at a given point with a certain energy, part of the energy of the incident photon is transmitted to the electrons in the medium. This electron is set in motion and then diffuses its energy during its journey through the medium. The distance traveled by this electron depends on its initial energy and the composition of the medium. The

absorbed dose is directly related to this energy transmitted locally by the electron. The latter can also lose its energy by the braking radiation. In this case, the energy lost by the electron does not participate in the energy absorbed. The braking radiation is made up of photons that will interact elsewhere in the medium. This implies that the energy transferred to one place in the medium is absorbed elsewhere.

I.5. Definition of volumes in radiotherapy

I.5.1. Volumes to be treated or target volumes

Advances in imaging and computer systems have made it possible to more clearly define the volumes of interest in radiotherapy. We will detail here the definitions of the volumes coming from the International Commission on. Radiation Units & Measurements report (ICRU 50, 1993). They are shown schematically in Figure 1.5.

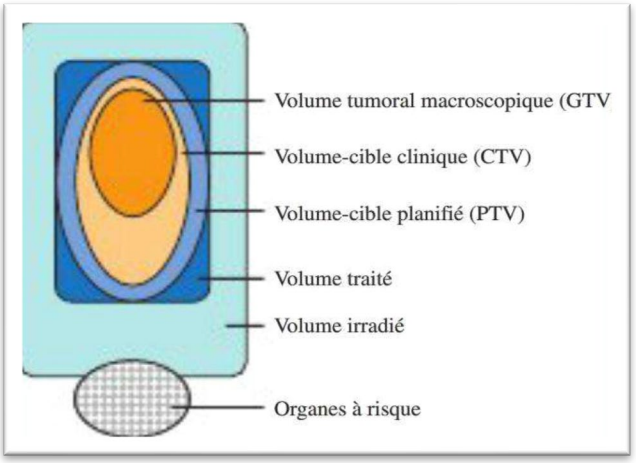


Fig.I.5. volumes of interest in radiotherapy.

Gross Tumor Volume: GTV

It is the one that is visible on the imaging (scanner, MRI). He will receive the stronger dose.

Clinical Target Volume (CTV)

It includes GTV, as well as tissues with a high tumor probability even if this is not visible on imaging. The definition of CTV is still subjective for many locations and is based on experience and knowledge of the disease (occult lymph node involvement, for example). The definition of GTV and CTV constitutes an essential part of the prescription.

Planning Target Volume: PTV

It includes the CTV and a safety margin that takes into account the positioning uncertainties, the possible movements of the organs and the patient.

The ICRU recommends optimizing the parameters of the treatment chain to homogenize the dose as much as possible inside the PTV. It is recommended that you plan so that the PTV dose is between 95% and 107% of the prescribed dose.

I.5.2. Volumes related to dose

The volume treated

This is the volume surrounded by an isodose surface specified by the radiotherapist, corresponding to a minimum dose level allowing the goal of treatment to be achieved. Ideally, this treated volume should correspond to the forecast volume (PTV).

The irradiated volume

It is the volume of tissues receiving a dose considered to be significant with respect to the tolerance of healthy tissues. The volume of the isodose corresponding to 80%, 50% or 25% of the prescribed dose can be evaluated, for example.

Volumes to protect

Organs At Risk (OAR) are tissues for which it is crucial to limit irradiation in order to limit side effects. Particular attention must be paid to the dose distribution to the OARs, mainly because of the importance of the gradients observed at the edge of the target volume. The dose constraints to OARs often intervene as penalties in the cost function to be optimized in order to define the treatment plan.

Three classes of organs at risk have been defined according to their level of morbidity:

- Severe morbidity: the organs, in the event of serious lesions, likely to cause total loss of function. For example, damage to the spinal cord making paraplegic, damage to the retina or optic nerves causing blindness etc.
- Moderate morbidity: organs whose lesion leads to significant functional loss. We find the salivary glands, the lens, the ears, etc.
- Transient morbidity: organs whose lesion leads to minor or no functional loss. For example, the skin or the mucous membranes.

The organization of the tissue is important in determining this morbidity:

- A series architecture corresponds to an organ with severe morbidity because the function depends on all its functional subunits. It can be represented by analogy with electronic circuits in series. The rupture of a single component results in the total loss of organ function. Overdose at one point of this organ therefore impairs the function of the entire organ. We are then interested in the maximum dose received by this tissue. This is the case with organs such as the spinal cord, the severing of which causes paraplegia downstream.
- A parallel architecture corresponds to an organ with moderate, even low or transient morbidity. The organ is made up of several functional subunits more or less independent of each other. Thus, the loss of organ function following irradiation requires the destruction of a significant number of subunits. If the volume destroyed by the irradiation is reduced, a repercussion on the organ and especially the quality of life of the patient is avoided. Thus, such an organ can receive a high dose if part of the volume is preserved. We are therefore interested in a constraint of the average dose or dose-volume type, that is, part of the volume must not be irradiated beyond a certain dose. One can quote like organ in parallel the parotids or the retina.

Finally, for each of the organs in series or in parallel, dose-volume relationships must be respected. This relationship can be represented by dose-volume histograms (DVH), which is explained next in section 10 of this chapter.

In practice, the notion of PTV has been extended to OARs. (ICRU 62, 1999) defined a planning volume for organs at risk (Planning risk volume, PRV). This volume corresponds to the volume of the OARs extended by a margin considering the movements or deformations of the OARs inside the body, as well as the consequences of the patient positioning uncertainties during the treatment. PRVs are preferably used for serial organs [4].

I.6. Radiation therapy techniques

I.6. 1.Conventional radiotherapy

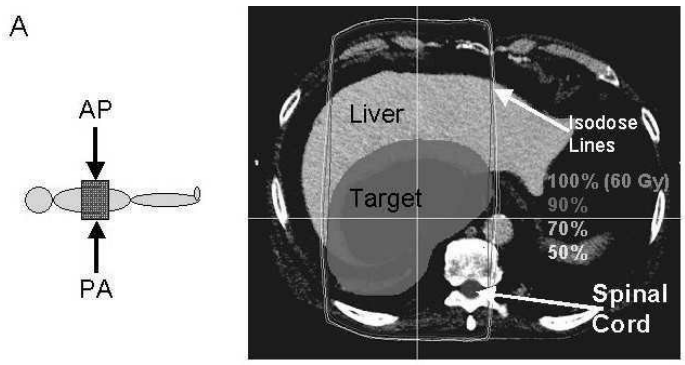


Fig.1.6. Example of isodoses in conventional radiotherapy.

In conventional radiotherapy, as a rule, the beams are wide enough to irradiate the entire target from the irradiation angles. Treatment planning was done "by hand", hence the use of beams with flat profiles and the use of 2 to 4 beams positioned on the 4 cardinal points in order to facilitate the already complex calculation. Figure 1.6 shows an example of treatment for liver cancer with two opposing parallel beams. The intersection of the two beams creates a high dose area near a rectangular shape that encompasses almost the entire irradiated volume of the patient. Unfortunately, this area contains a critical structure - the spinal cord. For this treatment to be viable, the dose prescribed by the doctor should be kept below the dose tolerated by the spinal cord. But the dose can then be wrong in the tumor. For years, this classical technique, which uses two to four beams for tumor treatment, has been the gold standard of radiation therapy.

The radiation goes via a device recognized as MultiLeaf Collimator (MLC), which is represented in Figure I.7, before reaching the patient. The MLC is perpendicular to the beam and consists of multiple movable metal components known as leaves that are used to block portion of the beam, allowing for more accurate form control. The leaves are arranged in pairs, one on top of the other. Different radiation approaches are described by how they use the LINAC, gantry, and MLC in combination. The most typical ones are listed below.

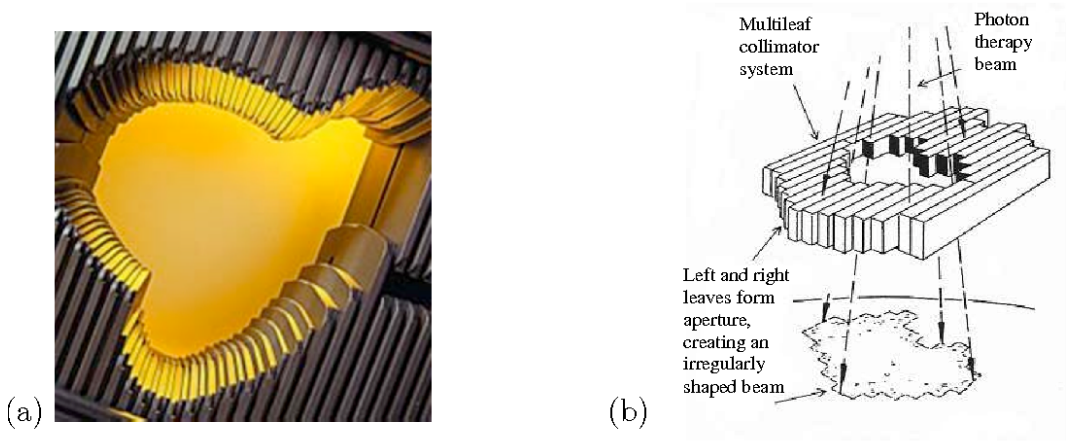


Fig.I.7. (a) A Multileaf collimator system; (b) An aperture projection to the desired configuration.

I. 6.2. Conformational radiotherapy

Conformational radiotherapy (3D-CRT) made its appearance at the end of the 1990s due to the advent of the computer and the development of the MLC. CT-scan allow 3D reconstructions of the body and all the organs. Beam Eye View (BEV) software, viewed from the target, allows virtual 3D treatment plans to be created that more precisely contour the tumor while sparing healthy tissue. This enables delivery of a dose distribution having a very high degree of conformance with the shape of the tumor.

These dose distributions are represented by what are called isodose curves. The latter illustrate the iso-levels of the absorbed dose. The level of isodose is defined as a percentage of the prescribed target dose. Figure 1.8 shows a dose distribution with different levels. The region with high dose is represented by the line 60 Gy (online black), which follows the shape of the tumor. The outer curve is 20% isodose, which means that the tissue inside this curve receives up to 20% of the prescribed dose. By delivering the highest dose according to the exact tumor shape, nearby healthy and critical tissue is spared.

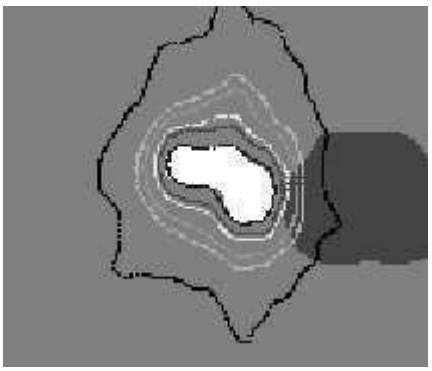


Fig.I.8. Example of isodoses in conformational radiotherapy.

Technically, to achieve this conformation, multi-leaf collimators are used to replace square fields. The most illustrative example is the treatment of prostate cancer. This 40 cm 3 organ has the shape of a pyramid. Before the end of the 1990s, it was irradiated by 8 cm square fields delivering a dose of 60Gy so as not to damage the rectum. These square fields irradiate a volume close to 500 cm 3 (half liter), i.e. approximately ten times more healthy tissue than tumor tissue. With the new 3D-CRT technique, the prostate is irradiated with beams reproducing the shape of the pyramid and no longer in the shape of a cube. The rectum is protected as well as the bladder, knowing that the dose in the target volume has been increased to 70-76 Gy. Randomized trials have shown that this dose increase is accompanied by an improvement in local control and survival. without an increase in rectal, bladder or even sexual complications. A high dose in a small volume remains a model of preference, subject to ballistic targeting accuracy. In 2000, 3D-CRT became the routine technique for the vast majority of irradiations, particularly for curative purposes (brain tumors, ear nose throat (ENT), lung, prostate, etc.).

Despite this therapeutic progress brought by 3D-CRT, it should be noted that this technique is less precise in the case where the tumor presents concave shapes with in addition organs at risk in these concavities, which represents 30% of cases. Hence the development of Intensity Modulation Radiation Therapy (IMRT) that we will present in the next section.

I.6.3. Intensity Modulated Radiation Therapy

The beam intensity can be adjusted (modulated) over the beam cross-section in IMRT, which is a generalization of 3D-CRT. This is done by superimposing many beams with different MLC configurations on top of each other. Each configuration is referred to as a beam segment, and the entire intensity distribution of the beam is made up of the superimposed intensities of all segments. Figure I.9 shows an illustration of an IMRT therapy.

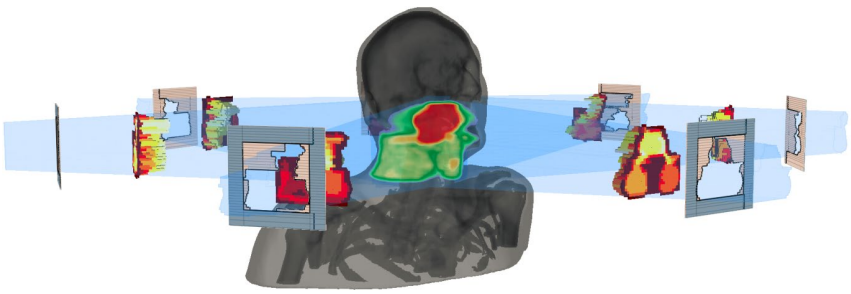


Fig.I.9. Illustration Through the superposition of many beams from different angles, an IMRT plan gives a focused dosage to the tumor volume. An SMLC plan is depicted here. The MLCs shape each beam individually, and the varied intensity across the beam is achieved via a series of shots with different MLC configurations.

In general, IMRT plans yield more concentrated dose distributions to the tumor than 3D-CRT plans [5], particularly when giving dose distributions that are concave or have steep gradients. However, because of the greater number of varied beam segments, IMRT plans normally take longer to complete. Longer delivery durations not only slow down the clinic's throughput (and hence lengthen waiting lists), but they also raise the chance of patient mobility during treatment, jeopardizing the plan's quality. Due to the MLC's failure to totally block off the beam during delivery, the patient is also exposed to some additional unwanted radiation. Finally, the restraining device utilized to restrict movement may cause discomfort to the patient during the fraction. However, in most circumstances, IMRT is considered to be a better option than 3D-CRT due to the potential to generate more conformal designs. In the United States, the percentage of radiation therapy treatments using IMRT increased from 0.15 percent to 95.6 percent between 2000 and 2008 [6]. A number of IMRT approaches have been developed, which are listed below.

I.6.3.1. Segmential MultiLeaf Collimation

The gantry and the MLC leaves are static during irradiation in Segmential MultiLeaf Collimation (SMLC), which is in some ways the simplest version of IMRT. During movement of either, the beam is turned off. Typically, a set of gantry angles is chosen first, followed by the creation of numerous distinct MLC shapes for each angle. Step-and-shoot IMRT is another name for SMLC. It's the most similar technology to 3D-CRT, but with a variety of MLC shapes for each angle.

I.6.3.2. Dynamic MultiLeaf Collimation

Allowing the leaves to move during the irradiation results in an extension of SMLC. This is known as Dynamic MultiLeaf Collimation (DMLC), and it has the benefit of shorter treatment times and more options for dosage shaping. However, because finite leaf velocities and accelerations (typically in the range of 1-4 cm/s and 50-70 cm/s2 [7]) must be considered, the treatment becomes more complicated. DMLC treatments often have longer beam-on times than SMLC treatments, but spend less time turning off the beam and adjusting the leaves, resulting in a lower net treatment time [8]. While this is a benefit of DMLC, the added complexity can be an issue, especially if a treatment must be discontinued and restarted. Sliding Window IMRT is another name for DMLC.

I.6.4. Volumetric Modulated Arc Therapy

During irradiation, the gantry is slowly rotated over the patient, resulting in Volumetric Modulated Arc Therapy (VMAT). An arc is the name for a continuous revolution. During rotation, the MLC leaves move so as to shape the beam intensity appropriate to the tumor. VMAT treatments offer the advantage of being able to deliver plans faster than SMLC and DMLC without sacrificing plan quality [9]. The additional complexity, like with DMLC, makes the treatments more difficult to manage. Nonetheless, VMAT is frequently seen as a viable therapy option.

1.7. Dosimetry and treatment planning

The use of radiotherapy as a means of treatment has arrived in clinics without the radiophysicists possessing powerful means of calculation. Therefore, the first treatment planning systems (TPS), were carried out using solutions based on empirical methods using a limited description of the patient's anatomy.

Major innovations then made it possible to refine the techniques for performing TPS. The anatomical description of patients has been developed with the emergence of new tools for medical imaging, such as CT scanners in the early 1970s, and more recently, the MRI, and the Single Photon and Positron Emission Tomography (SPECT and PET) scanners.

On the other hand, the increase in the computation capacity made it possible to develop numerically calculated treatment planning since 1970s, limited to simple geometries without considering heterogeneities. Subsequently, improvements were made such as the heterogeneous zones calculation in 1983, using fine discretization. At the same time, these technical innovations have allowed a refinement of simulation techniques.

I.8. Treatment Planning

The treatment plan is the whole set of instructions supplied to the equipment to perform the treatment. This covers all beams, their orientations, MLC leaf positions, and each MLC configuration's radiated power. The treatment planner's ultimate goal is to select a plan that promotes long-term tumor control without creating difficulties to healthy tissue.

The patient's 3D image collection is provided by CT imaging. The clinician uses this image set to outline the contours of the tumor and important healthy organs nearby. This combined dataset (CT and contours) is then sent to a treatment planner, who chooses beam angles and

works on optimizing a treatment plan, as shown in figure 1.10. This is the step that we suggest our contribution in this paper.

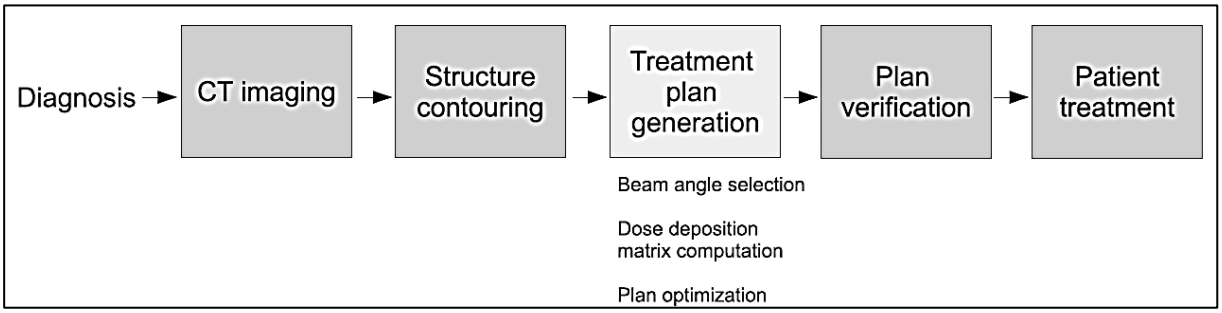


Fig.I.10. Basic Radiotherapy treatment planning workflow.

To produce a deliverable treatment plan, the optimized fluence levels must be converted to multi-leaf collimator locations and monitor units in the actual clinical process. At this stage, the deliverable's treatment plan is examined., and it is used to treat the patient once it passes this quality assurance (QA) step, whereas, the plan is delivered to a device recognized as a phantom.

The phantom is usually a body of water with a sensor grid for measuring the radiation supplied. The plan can be administered to the patient if the difference between the calculated and delivered doses is less than a certain threshold.

I.8.1. Treatment planning systems

The heart of RT systems and the key to better patient outcomes are treatment planning systems (TPS). Following the loading of picture files and the identification of tumors, the systems create a sophisticated plan for each beam line path for how the therapy system will deliver radiation. Figure I.11 depicts an open source TPS in action.

The software also calculates the projected dose distribution in the patient's tissue, taking into account factors like tissue energy level penetration and the type of tissue that the beam lines pass through (e.g., bone or lung vs. muscle). These devices also assist in beam placement by avoiding important structures that are more vulnerable to radiation in order to decrease therapy-related collateral harm. This could include complex automated programming for MLC leaf sequencing to shape the beam around important structures during dose administration. These treatment plans can also be tweaked to account for tumor shrinkage over the course of treatment.

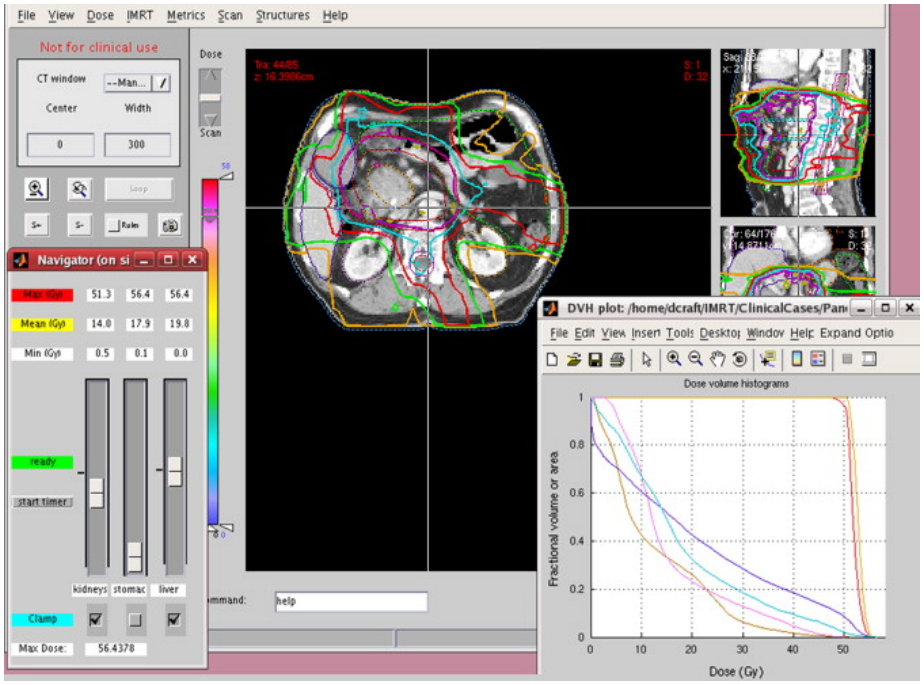


Fig.I.11. A computational Environment for Radiotherapy Research (CERR) window which is a MATLAB based software platform for developing and sharing research results using radiation therapy treatment planning and imaging informatics [10].

I.9. Dose Calculation methods: a state of the art

In radiotherapy, it is essential to have a precise knowledge of the dose delivered to the target volume and to the neighboring critical organs. Today, the dose is calculated in three dimensions using algorithms implemented on TPS that take into account the anatomical characteristics of the patients and the physical and geometric characteristics of the beams.

Typically, dose calculation algorithms model the physical energy transfer processes and calculate the dose deposited by treatment beams into the voxels that make up the patient's digital phantom. The dose received by the patient is divided into two categories: first, the dose from primary photons which have not interacted in the head of the accelerator. These photons only interact with the patient's tissues, causing dose deposition by moving electrons and (indirect) dose deposition by scattered photons in the phantom. Second, the dose from the primary photons that interacted with the accelerator head produces contaminating electrons and scattered photons from the accelerator head [11]. Contaminated electrons deposit their energy on the patient's skin. Photons scattered from the head of the accelerator, in turn, deposit their energy in the patient.

These dose calculation algorithms for complex situations (considering heterogeneities, surface irregularities, etc.) are sometimes very imprecise. They give an approximate representation of the dose distribution in the patient. The level of approximation will be different depending on the type of algorithm used.

The point is, it is inevitable to make tradeoffs between accuracy and speed of computation. To date, the algorithms used are more and more sophisticated and aim at improving the precision in the calculation of the dose. The commonly used dose calculation models are briefly described and commented on here.

I.9.1. Monte Carlo method

Monte Carlo algorithms (MC) are stochastic methods for solving numerical problems for which no analytical formulation can be obtained. In the case described here, the physical models of the transport and diffusion process of electrons and photons are in fact statistical models of radiation-matter interaction at the particle scale (photons, electron, atom). The complete analytical formulation of such a model is impossible, so the Monte Carlo method [12,13] is applied to it. But by extension, in dose calculation, the statistical model and stochastic resolution duo are often called “Monte-Carlo method”.

The Monte Carlo method has very often and very completely been detailed and commented on in the literature. It is based on the simulation of the transport of particles from their production to the deposition of energy in matter. It is the most realistic current method. The dose distributions obtained by the Monte Carlo method are therefore used as references.

On the other hand, to obtain precise results with an acceptable uncertainty, it is essential to simulate the transport of a large number of particles (at least of the order of 107). This implies relatively long computation times which can reach several days [14,15].Many methods have been developed to speed up the dose calculation and ultimately allow the use of this approach in clinical routine. To date, the Corvus v.5 (NOMOS Corporation, Swickley, PA) TPS has adopted the Monte Carlo method for the dose calculation. Despite everything, this method is still not very suitable for routine digital dosimetry. They are mainly reserved for theoretical studies, in particular to test the validity of other models in complex situations, difficult to achieve on the experimental level and appear in the clinical validation process for complex and critical cases.

I.9.2. Methods based on the separation of primary and scattered radiation

This method was used for a long time in TPS during the 90s. This method was first developed by Clarkson in 1941 [16], then by Cunningham in 1972 [17,18]. This method consists in calculating separately the "primary dose" and the "diffused dose". The dose at a point is the sum of the contributions of the primary and scattered components of the radiation field. The detailed presentation can be found in [15]. This method is particularly suitable for conformational radiotherapy. Nowadays, it is still used in Cyberknife's dedicated TPS, but less used in IMRT.

I.9.3. Convolution / superposition methods

The convolution / superposition methods, proposed in the early 1980s [19-22], separate the processes of transport and deposition of energy in two phases: transport of energy by primary photons, and its deposition by secondary particles (electrons and photons). In reality, electrons from photoelectric, Compton and pair creation interactions deposit their energy within a few millimeters or even centimeters around, but Compton interactions also release secondary photons which can travel long distances before interacting again and depositing all or part of their energy [23]. The deposition of energy by these secondary particles is also integrated into the expression of Kernels. The total absorbed dose at a given point results from the sum of the dose distributions calculated from all the points where energy is released (superposition), this sum is achieved by a convolution of these two phases, illustrated in Figure 1.12.

The general dose calculation is written:

$$Dose\ (Q) = \sum_{P} T(P) \times K(P - Q)$$

Where: *T(P)* is the Terma at point P, *K (P - Q)* gives the share of the energy *T(P)* deposited at point *Q*.

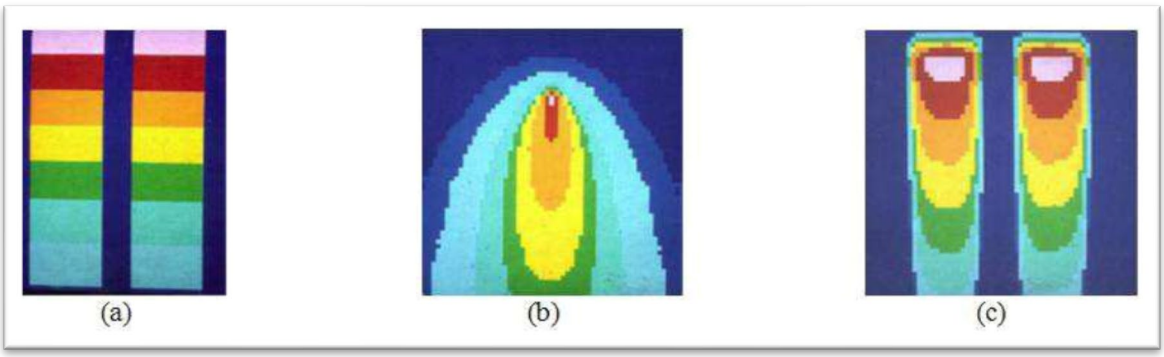


Fig.1.12. (a) Primary photon fluence (b) Convolution nucleus (c) Dose distribution

I.9.3.1. Kernel point method

The kernel point method as shown in Figure 1.13 (a), consists of two successive phases. The first phase corresponds to calculating the total energy transferred by all the primary photons in a unit of mass (Terma). The second phase is to calculate a deposition model of this energy around a primary interaction site.

I.9.3.2. Pencil beam method

A pre-integration of all the nuclei in the direction of the infinitesimal section beam depth makes it possible to obtain the absorbed dose due to the latter. Such a nucleus is shown in Figure 1.13 (b). This so-called pencil beam approach is used for electron beams [24] and photon beams [25].

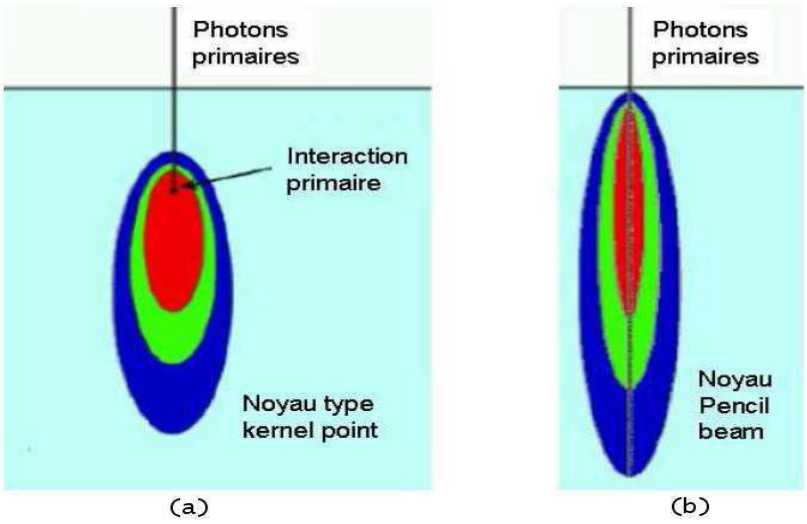


Fig.I.13. Convolutional nuclei: (a) nucleus representing the distribution of energy released by interactions of primary photons taking place at a single point; (b) a core of the pencil beam type, representing the dose deposited by a beam of infinitesimal section.

I.9.3.3. Collapsed cone convolution method

The so-called "Collapsed cone convolution" method, introduced by Ahnesjo in 1989 [21] offers one of the best time / precision compromises.

For a voxel with a given Terma, the method considers that the energy transport is done according to cones in the different directions and starting from the central point (the point where the Terma was previously calculated) (see Figure 1.14). Then, it assumes that all the energy that is propagated in a cone is transported, attenuated and deposited according to an exponential law on the axis of that cone.

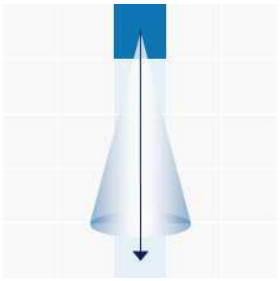


Fig.I.14. A Terma voxel (blue cube) emits a cone of energy. The dose is deposited only in the voxels which are on the axis of this cone (arrow).

This assumption makes it possible to simultaneously distribute the dose along a set of discrete lines that emerge from each Terma point and to accumulate the Terma as one advances in the direction considered (see Figure 1.15). This simultaneity is the basis of the algorithm's performance in computing time.

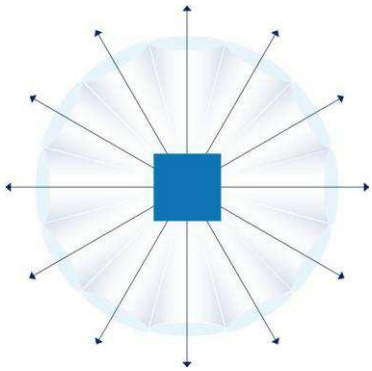


Fig.I.15. The propagation of the energy of a Terma (cube) voxel along the cones

The resulting method is today one of the most efficient, as it offers acceptable precision (rarely more than 5% error [26,27]), while taking a limited time (of the order of a minute). But this time is still too long to consider iterative optimization of treatment plans.

I.9.4. Calculation methods by neural networks

In recent years, various authors have used neural networks, taking advantage of their properties of universal approximators [28] to quickly and precisely calculate the dose. It has been shown that a neural network is able to "learn" the dose in a homogeneous medium on the beam axis [29] and the dose on a 2D plane in homogeneous volumes [30]. The most sophisticated approach to date, called NEURAD [31,32] is capable of calculating the dose absorbed in heterogeneous medium by a wide beam from doses in medium homogeneous previously learned by neural networks. This method is based on the following two assumptions:

1. In a homogeneous sub-part of a heterogeneous phantom, it is assumed that the dose can be obtained from a pre-calculated model in a homogeneous medium, and learned by a neural network.
2. Electronic equilibrium is assumed to be achieved in the vicinity of interfaces (interfaces crossed by all or part of the primary beam). The consequence of this assumption is that the dose does not undergo strong variations near these interfaces.

This method calculates the dose for an interface orthogonal to the beam axis. It considers that there is equality between the dose at the last point of the first material and the dose at the first point of the second. They use the dose curves, obtained via their neural network, in the first and the second material. The dose before the interface is considered equal to the dose in a homogeneous medium. The dose after interface is too, but it is taken from the depth where it is equal to the dose at the last point before the interface. This way, the dose curve at the interface is very continuous.

The results of this method are very correct in the case of a wide beam. On the other hand, the case of narrow beams is not correctly treated (in fact the method exploits the theorem of Fano [33] which is not valid by hypothesis in the case of narrow beams). A method of processing interfaces parallel to the axis of the beam, and within it, is also given [31,32]. However, the issue of oblique interfaces, as well as that of interfaces outside the bundle, is not addressed. The calculation time on 3D grids is greater than a minute.

Despite the limitations stated above, the method contributed to the idea of using, in a homogeneous subpart of the phantom, a pre-calculated dose distribution in a homogeneous phantom made of the same material. The major issue is to manage the electronic imbalance near the interfaces in the case of small beams. But another important point is also to be considered in order to have the best performance.

The methods presented so far calculate complete dose distributions. In other words, if one wishes to calculate the dose only at a single position, it is necessary to make numerous intermediate calculations on other points, or even to make a complete dose calculation on a large part of the ghost.

However, obtaining a complete dose delivery on a complete phantom is not always necessary. For example, if the dose must be known in a complete and precise manner on the tumor and the organs at risk, on the other hand, it is possible to reduce the density of checkpoints

on the less sensitive areas. Another example is the optimization of the treatment plan, for which it may be sufficient, especially in the preliminary stages of placement and orientation of the beams, to have only the dose available at certain points. For example, a few control points may be enough to verify if the orientation of a beam is correct, or if it should be changed.

In our work, we have used a pencil beam dose calculation algorithm referred to as the quadrant infinite beam (QIB) model due to its availability on (CERR) [25,34] (see Chapter IV).

I.10. Quality criteria in radiotherapy

The goal of radiotherapy has always been to deliver the prescribed dose value to the tumor and the minimum dose to related areas. There are several ways to assess the quality of the treatment plan depending on the type of representation used.

I.10.1. Isodose Curves

The first way to represent a dose distribution is to use a plane representation of the isodose curves. An isodose curve, shown in the figure 1.16, represents a set of points of the irradiated medium where the value of the deposited dose has the same value. This means of control can be done quickly visually, cut by cut. But it has the disadvantage of making the comparison between several treatment plans difficult and imprecise.

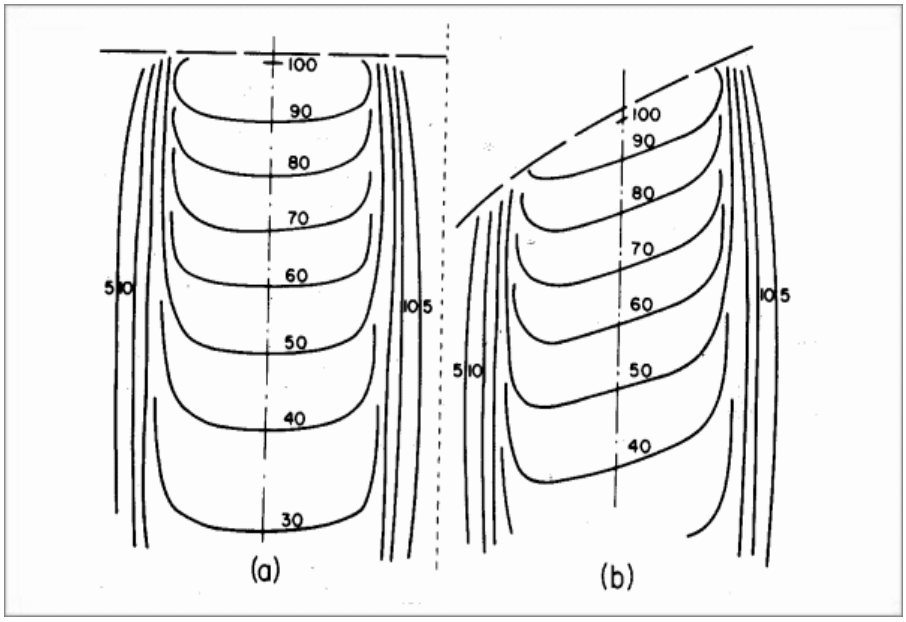


Fig.I.16. Normal (a) and oblique (b) incidence isodose curve

I.10.2. Dose Volume Histogram

To allow better visualization and study of these yields, another quantification of the deposited dose is implemented in the form of a dose-volume histogram (DVH) [3]. This representation, allows to know the detail on the doses deposited in each volume of interest, for example, targets, critical structures, etc. A DVH not only provides quantitative information with regard to how much dose is absorbed in how much volume, but also summarizes the entire dose distribution into a single curve for each anatomic structure of interest. It is, therefore, a great tool for evaluating a given plan or comparing competing plans.

The DVH may be represented in two forms: the cumulative integral DVH and the differential DVH. The cumulative DVH is a plot of the volume of a given structure receiving a certain dose or higher as a function of dose (Figure 1.13). Any point on the cumulative DVH curve shows the volume that receives the indicated dose or higher.

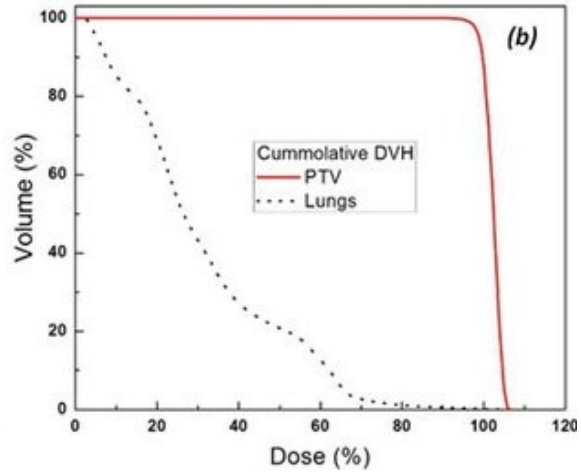


Fig.I.17. A cumulative DVH from a radiotherapy plan.

The differential DVH is a plot of volume receiving a dose within a specified dose interval (or dose bin) as a function of dose.

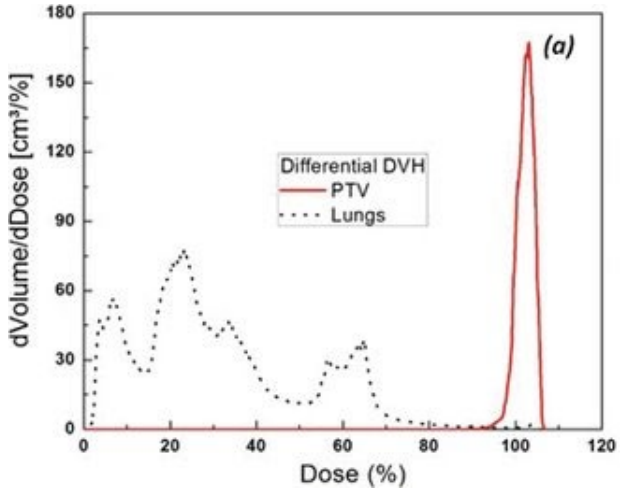


Fig.I.18. A Differential DVH from a radiotherapy plan.

The differential form of DVH, as shown in Figure 1.18, depicts the magnitude of dosage variation within a given structure. A single bar of 100 percent volume at the indicated dose, for example, is the differential DVH of a uniformly irradiated object. The cumulative DVH has been proven to be more beneficial and is utilized more frequently than the differential form of DVH. All of these comparisons are based on the analysis of the relationships between the different parameters characterizing the treatment.

I.11. Optimization in IMRT

Optimization has usually been divided into three sub-problems, each of which focuses on fixing a subset of these variables and optimizing a subset of other parameters. Number and beam angles optimization (BAO), fluence map optimization (FMO), and MLC segmentation optimization (MLCSO) are the three sub-problem categories as summarized in figure 1.19. The FMO is the core of this dissertation's IMRT-based optimization problem; nevertheless, a brief review of the BAO and MLCSO will be offered, followed by a discussion of the FMO.

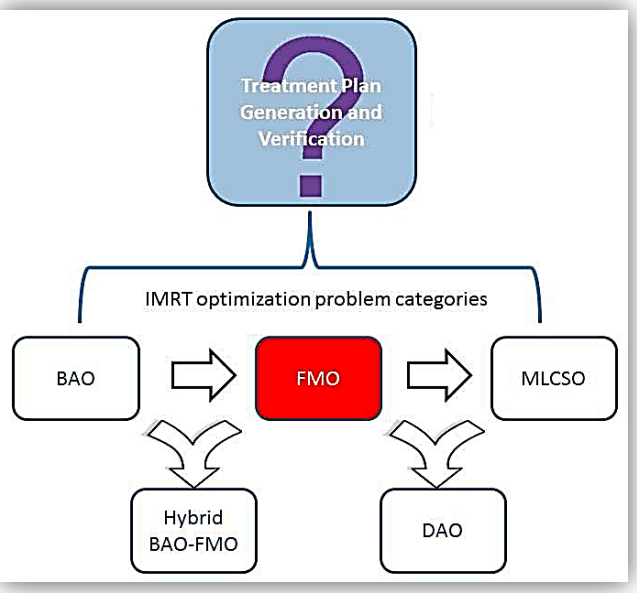


Fig.I.19. The main IMRT optimization challenges: BAO, FMO, MLCSO, and Hybrid BAO-FMO. The FMO is marked in red since it is the IMRT optimization challenge that was discussed in this thesis.

I.11.1. The problem of Beam Angle Optimization (BAO)

Prior to IMRT, many plans were made up of a small number of manually chosen beams, like two parallel-opposed beams for breast treatments or four-field box for prostate treatments. With the introduction of IMRT, it became possible to use additional beams to better spread out the dose to healthy structures. When working with a limited number of beams, the location of

these beams is essential, especially in non-convex geometries where the tumor is wrapped around or near critical organs. Increasing the number of beams in a treatment plan will lengthen treatment duration and, as a result, increase the risk of an undesirable outcome due to patient movements. As a result, determining the ideal number and orientation of treatment fields for a given patient geometry and set of dosage limitations in IMRT entails determining the optimal number and orientation of treatment fields.

The problem is characterized as picking a certain number of beams, b, from a bigger set of all possible beams, k, that will result in the best objective function value, f(x), from a larger set of all possible beams, k [35-36]. The objective function is expressed as a function of the vector x. Each element of this latter is a fluence value allocated to a specific beamlet of the chosen beam. The dealing between dosage and fluence is then utilized to frame the objective function in terms of one of the RT treatment goals, such as minimizing the mean dose to a crucial structure. Due to the risk of collisions between the patient and the LINAC, this problem has traditionally been limited to only coplanar beams. To keep track of the beams included and avoided in the solution from the set of k potential beams, a binary beam selection variable is necessary. As a result, mixed-integer optimization techniques for the BAO issue have been a natural choice [37,38]. Single-step procedures, which include vector quantization and scoring methods; and set scoring models and alternative iterative techniques, which include local search methods, simulated annealing, and evolutionary algorithms, are two ways to classify the various approaches taken for the BAO problem. In addition, researchers have looked into merging the BAO and FMO to determine the ideal number of beams, direction, and fluence patterns all at once [39]. Despite the fact that there have been numerous publications on the subject of BAO and that it remains a hot issue of research, the majority of institutions that use IMRT still use a trial-and-error manual approach to picking beams for each treatment plan. Prostate cancer cases, in general, allow for the use of a typical coplanar beam arrangement template in terms of number and orientation for the majority of prostate instances. More complex instances, such as head and neck, necessitate the creation of a patient-specific beam arrangement.

The next logical step in the IMRT optimization process is to establish the ideal fluence pattern for each beam that will result in the best-possible dose distribution in regard to some preset prescription dose and dose constraints to the relevant structures. The fluence map optimization (FMO) challenge is based on determining these ideal fluence patterns for a specific set of beams. The FMO is discussed at the end of this subsection because it is the foundation for the IMRT-based optimization problem studied in this thesis.

I.11.2. The problem of MLC Segmentation Optimization (MLCSO)

After determining the ideal fluence patterns using the FMO, it's time to figure out how to recreate this intensity distribution when treating the patient. Multi-leaf collimators (MLCs), tomotherapy, and compensator-based techniques have all been presented as standard approaches for delivering intensity modulated fields of IMRT. Despite the fact that the latter strategy is the least ideal since it requires more treatment time to swap compensators for each field, investigations have proven that it is a viable method for delivering IMRT fields [40]. Tomotherapy is a revolutionary approach of treating patients with intensity modulated radiation treatment (IMRT), in which IMRT beams are administered slice by slice, similar to how CT imaging is done [41]. The MLC-based approach, on the other hand, is currently the most prevalent method for IMRT treatments. The MLC allows you to realize the complex intensity distributions that the FMO problem generates. However, constructing MLC sequences for an arbitrary fluence map is not a simple task, and there are some circumstances where the fluence distributions generated by the FMO are physically impossible to actualize in the MLCSO problem. As a result, the MLCSO's overall objective is to provide a collection of deliverable MLC sequences that produce (sequenced) intensity distributions that are as near as feasible to the optimum (reference) fluence maps produced by the FMO issue.

As stated in section 6, MLCs may be utilized in three different ways to materialize and execute IMRT treatment plans: dynamic mode, static mode, and arc therapy. Many variants on the MLCSO issue have been proposed, trying to include various physical phenomena that occur when MLCs are used, such as the tongue and groove effect, dynamic leaf gap, intra-leaf radiation transmission, and so on. As previously stated, there are certain difficulties in recreating the optimum fluence map (gained from the FMO) while solving the MLCSO, since the FMO does not always take into account the physical and mechanical constraints of the delivery mechanism - the MLCs. As a result, one novel method to resolving this challenge is to combine the FMO and MLCSO issues, with a predefined set of MLC sequences that are known to be deliverable being utilized as input in calculating the optimum fluence. As a result, determining the optimum weighting system in terms of fluence for each of these deliverable MLC segments is basically the goal. DAO (direct aperture optimization) is the name given to this kind of optimization issue [42,43].

I.11.3. The problem of Fluence Map Optimization (FMO)

The BAO and the MLCSO are important and well-studied RT optimization problems, but the FMO problem is the focus of this thesis since it is the most natural point in the IMRT optimization process to analyze a specific beam parameter that has not been investigated directly. The beamlet energy used in IMRT beams is this exact beam parameter. An overview of the FMO will be presented before exploring the possible benefits of implementing beamlet energy optimization into the FMO.

There has been a lot of research focused on addressing and enhancing the FMO since the advent of IMRT and inverse planning in the 1980s. Webb and Bortfeld used different optimization methods to frame the issue, but they both had the same goal in mind: to find the set of optimal fluence maps given some specified geometry and a desired dose distribution. The objective function and constraints, as previously stated, are the essential components of an optimization problem, both of which are defined in terms of the decision variable for which an optimal set of values is sought. Whatever optimization method may be used to solve that specific formulation of the optimization issue depends on whether or not constraints and/or an objective function are included, as well as which form (e.g. linear, non-linear) these functions are pursued. Different optimization models and algorithms (e.g. stochastic vs deterministic) may be classified in a number of ways, especially when applied to the FMO. Ehrgott et al. [44], for example, divide the FMO into five categories: feasibility, linear programming, non-linear programming, mixed-integer programming, and multi-criteria optimization.

I.11.3.1. relationship between Dose and fluence

Before going through each one in detail, it's important to note that the relationship between dose and fluence is a feature that all algorithms and models for the FMO share. This connection is also necessary for MLCSO, BAO, and any other IMRT parameter optimization study. The relationship between intensity and dose has previously been shown to be roughly linear [45]. We are dealing with a predefined beam arrangement for the FMO, so let's assume beams $\boldsymbol{b}$ = 1, 2,..., $\boldsymbol{B}$ contribute to the beam arrangement in this example. Each beam can be subdivided into smaller beamlets, with $\boldsymbol{j}$ = 1, 2,..., $\boldsymbol{N}$ in this scenario. Finally, the volume in question will be divided into smaller rectangular sub-volumes known as voxels, with indices i = 1, 2,..., $\boldsymbol{M}$ describing these voxels as shown in figure I.20.

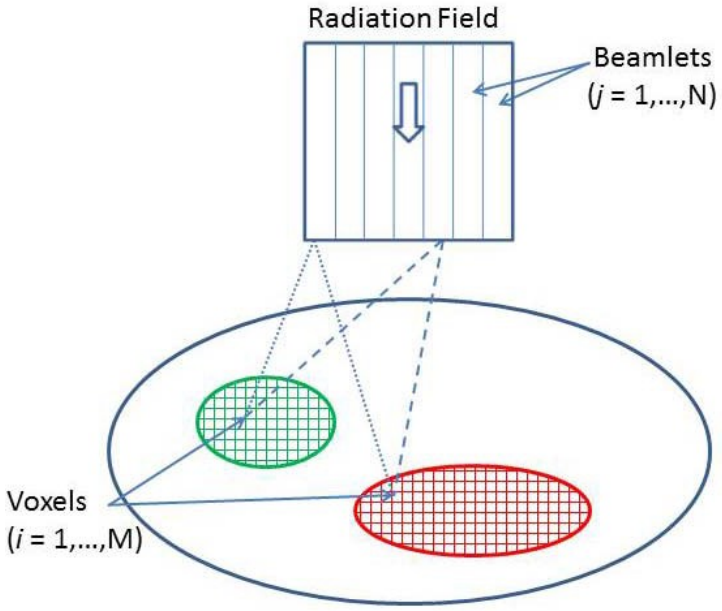


Fig.I.5. Radiation field discretization into beamlets and volume of interest into voxels on a 2D transverse slice. Multiple beamlets contributing fluence to the same voxel are depicted..

The dose deposited to voxel ***i*** by unit intensity of beamlet ***j*** from beam ***b*** is then commonly represented by an element a_{ijb} which is a pre-calculated value. Alternatively, the beam and beamlet information can be combined into a single index ***k***, which contains an index for all beamlets. For example, if we have two beams, each with ten beamlets, then ***k*** =1,2,...,10, 11,...,20 for the total number of beamlets from all beams. In this scenario, a_{ijb} becomes a_{ik}, and all a_{ik} values can be sorted into a huge matrix known as the dosage deposition matrix, which is indicated by ***A*** or ***D*** of dimensions ***M*×*K***. The voxels are represented by the ***M*** rows, and the beamlets are represented by the ***K*** columns in this matrix. This dose deposition matrix can be calculated using any of the dose calculation methods outlined in section 9 earlier. The dose deposition matrix, which is described by the following linear equation, completes the linear relationship between dose and fluence:

$$\boldsymbol{d = Ax} \qquad \text{(I.1)}$$

Where ***d*** is a $\boldsymbol{M \times 1}$ dose vector with each element d_i representing the dose deposited to voxel ***i***, ***A*** is the $\boldsymbol{M \times K}$ dose deposition matrix, and ***x*** is the $\boldsymbol{K \times 1}$ fluence vector with each element x_k representing the fluence of beamlet ***k*** that one is seeking to discover optimal values for in the FMO issue. Equation I.1 is the foundation of any FMO problem, and it is further examined in Chapter IV in terms of include energy optimization in the FMO.

I.11.3.2. The feasibility method

The feasibility approach is not an optimization strategy, according to the five FMO model and algorithm classes, because it does not use an objective function that is minimized or maximized to discover the best solution. Instead, it depends entirely on dosimetric constraints to define the problem's feasible zone and assure that any solution picked from this region meets the desired constraints, regardless of which solution is deemed "optimal." This was one of the first efforts to using the FMO to produce acceptable fluence maps, and the groundwork for it was laid by using Cimmino's simultaneous projection algorithm, which projects the current iteration simultaneously onto all associated sets and determines the centroid.

I.11.3.3. Non-Linear Programming (NLP)

When it comes to commercial treatment planning systems, non-linear programming (NLP) models and algorithms have perhaps been the most popular for addressing the FMO [46,47].

One of the reasons is that the weighted least squares model has a simple interpretation and implementation when it comes to clinical restrictions, which are defined as the intended dose to various structures of interest (such as tumors (T) and organs at risk (OARs). This algorithm can be expressed in a number of ways, the most basic of which is

$$\min_{s.t\ \ x\geq 0} \frac{w_T}{M_T}\|A_T x - T_{PD}\|^2 + \frac{w_{OAR}}{M_{OAR}}\|A_{OAR}x - UB_{OAR}\|^2 \qquad \text{(I.2)}$$

All-important variables from Equation I.2 are listed in Table I.2. Bortfeld [48] demonstrated that there are no local minima for simple least squares objective functions like the one in Equation I.2. However, when dose-volume based limitations are included, this is no longer the case.

Variable	Description
A_T, A_{OAR}	rows of the dose deposition matrix A that correspond to the tumor (T) and OAR voxels, respectively
M_T, M_{OAR}	The total number of voxels associated with tumor and OAR structures, respectively.
T_{PD}, UB_{OAR}	T_{PD} = Prescribed dose of tumor. UB_{OAR} = upper bound for OAR dose
w_T, w_{OAR}	Priority weighting factor for tumor and OAR

Table.I.1. Least squares optimization model variables description (Equation I.2)

I.11.3.4. Mixed-Integer Programming (MIP)

Due to its ability to manage dose-volume restrictions, mixed-integer programming (MIP) provides an alternate way to solve the FMO. To summarize, MIP provides a binary variable that allows the algorithm to decide whether or not a voxel can exceed the prescribed dose level for the structure in which it is contained. The sum of the voxels that are allowed to exceed the dose prescription must be less than or equal to the dose-volume constraint's prescribed volume. The main disadvantage of adopting MIP-based techniques for FMO is that these programs are computationally costly due to the addition of thousands of variables for each dose-volume restriction simulated. This has been a major issue in commercial solvers' reluctance to use MIP-based techniques for the FMO problem.

I.11.3.5. Linear Programming (LP)

Linear programming is a method of optimization in which the objective function and constraints are both linear functions. George Dantzig created the Simplex method, which was an effective approach of tackling LP issues, in 1947, and developed the broad formulation of linear programming used for planning United States air force related-projects in 1946. The first linear optimization model for RT treatment was created in 1968 [49]. There have been various distinct applications of LP to RT since then. Shepard et al. [50] offered an outline of the use of LP to the FMO problem from the perspective of IMRT.

I.11.3.6. Multiobjective Programming (MOP)

The Multiobjective Programming (MOP) methodology has been discussed last because it was the method used to define the IMRT-related optimization problem that took up the majority of this dissertation. For the FMO problem, multi-objective or multi-criteria programming (MOP) approaches have lately gained popularity. The objective function in both the NLP and MIP techniques gives a single value as a representation of the solution quality. This is counterintuitive since the objective function is usually made up of many, competing sub-objectives, such as the dose to the tumor and the dose to the OARs. Rather than providing a single solution linked with a single optimal objective function score, the MOP technique provides an entire sequence of solutions, all of which fall on the Pareto front or non-dominated solution set. The tradeoffs between opposing sub-objectives can then be investigated, and a solution that best fits the physician's needs can be chosen from this non-dominated collection of alternatives. The weighted sum method, which may be used to obtain a non-dominated set of answers by employing multiple weighting factors and a genetic algorithm approach to solve the weighted sum issue, is the most frequent method for MOP. In Chapter III, the specifics will be discussed.

I.12. Conclusion

A state of the art in radiation therapy was detailed in this chapter, in terms of a brief RT history, radiobiology, basis notion, modes delivery techniques, quality assessment and optimization problems in order better understanding what is coming next in this thesis.

Referenes

[1] Anders Ahnesjo and Maria Mania Aspradakiso. Dose calculations for external photon beams in radiotherapy. Physics in Medicine and Biology, 1999.

[2] E Mladenov, S Magin, A Soni, G Iliakis, DNA double-strand break repair as determinant of cellular radiosensitivity to killing and target in radiation therapy, Frontiers in oncology, 2013.

[3] R.E. Drzymala, R. Mohan, L. Brewster, J. Chu, M. Goitein, W. Harms, M. Urie, Dose-volume histograms, International Journal of Radiation Oncology*Biology*Physics, Volume 21, Issue 1, 1991.

[4] A McKenzie, M van Herk, B Mijnheer, Margins for geometric uncertainty around organs at risk in radiotherapy, Radiotherapy and Oncology, 2002.

[5] J Purdy, Dose to normal tissues outside the radiation therapy patient's treated volume: a review of different radiation therapy techniques, Health Phys, 95(5):666-676, 2008.

[6] N Sheets, Intensity-modulated radiation therapy, proton therapy, or conformal radiation therapy and morbidity and disease control in localized prostate cancer, JAMA-J, Am. Med. Assoc, 307(15):1611-1620, 2012.

[7] K Wijesooriya, C Bartee, JV Siebers, SS Vedam, and PJ Keall. Determination of maximum leaf velocity and acceleration of a dynamic multileaf collimator: Implica- tions for 4d radiotherapy. Medical physics, 32(4):932–941, 2005.

[8] P Alaei, P D Higgins, R Weaver, and N Nguyen. Comparison of dynamic and step- and-shoot intensity-modulated radiation therapy planning and delivery. Medical Dosimetry, 29(1):1–6, 2004.

[9] M Teoh, C Clark, K Wood, S Whitaker, and A Nisbet. Volumetric modulated arc therapy: a review of current literature and clinical use in practice. Brit. J. Radiol., 84(1007):967–996, 2011.

[10] JO Deasy, AI Blanco, VH Clark , CERR: a computational environment for radiotherapy research, Medical physics, 2003.

[11] Ahnesjo A, Aspradakis M.M Dose calculations for external photon beams in radiotherapy [Revue] // Phys Med Biol. - 1999. - Vol. 44. - pp. 99-155.

[12] Andreo P, Monte carlo techniques in medical radiation physics [Revue] // Phys. Med. Biol.. - 1991. - Vol. 36. - pp. 861–920.

[13] Salvat F, Practical aspects of monte carlo simulation of charged particle transport : Mixed algorithms and variance reduction techniques [Revue] // Radiat. Environ. Biophys. - 1998. - Vol. 38. - pp. 15-22.

[14] Blanpain B, Mercier D, The delta envelope: A technique for dose distribution comparison [Revue] // Med.Phys.. - 2009. - pp. 797-808.

[15] Menguy Y, Optimisation quadratique et géométrique de problèmes de dosimétrie inverse // Thèse : mathématique appliquées. - Grenoble : [s.n.], 1996.

[16] Clarkson J.R, A note on depth doses in fields of irregular shape [Revue] // Brit. J. Radiol. - 1941. - Vol. 14. - pp. 265 - 268.

[17] Cunningham J. R Scatter-air ratios [Revue] // Phys. Med. Biol. - 1972. - 1 : Vol. 17 . - pp. 42-61.

[18] Bjarngard B. E, Rashid H, Obcemea C. H, Separation of primary and scatter components of measured photon beam data [Revue] // Phys. Med. Biol. - 1989. - 12 : Vol. 34 . - pp. 1939-1945.

[19] Boyer. A, Mok. E.A, photon dose distribution model employing convolution calculations [Revue] // Med Phys. - 1985. - Vol. 12. - pp. 169-177.

[20] Mohan R, Chui C, Lidofsky L, Differential pencil beam dose computation model for photons [Revue] // Med Phys. - 1986. - Vol. 13. - pp. 64-73.

[21] Ahnesjo A, Collapsed cone convolution of radiant energy for photon dose calculation in heterogeneous media [Revue] // Med Phys. - 1989. - Vol. 16. - pp. 577-592.

[22] Mackie T, Scrimger J, Battista J, A convolution method of calculating dose for 15-mv x rays [Revue] // Med Phys. - 1985. - Vol. 12. - pp. 188-196.

[23] Childress Nathan, Dose Calculation Algorithm [Revue] // Mobius3D White Paper. - 2012.

[24] Kooy H.M, Rashid H, A three-dimensional electron pencil-beam algorithm [Revue] // Phys. Med. Biol. - 1989. - Vol. 34. - pp. 229–243.

[25] Ahnesjo A, Saxner M, Trepp A, A pencil beam model for photon dose calculation [Revue] // Med. Phys. - 1992. - Vol. 19. - pp. 263–273.

[26] Vanderstraeten B, Reynaert N, Paelinck L, Accuracy of patient dose calculation for lung IMRT: A comparison of monte carlo, convolution/superposition, and pencil beam computations [Revue] // Med. Phys. - 2006. - Vol. 33. - pp. 3149–3158.

[27] Fogliata A, On the dosimetric behaviour of photon dose calculation algorithms in the presence of simple geometric heterogeneities : comparison with monte carlo calculations [Revue] // Phys. Med. Biol. - 2007. - Vol. 52. - pp. 1363–1385.

[28] Dreyfus G, Réseaux de neurones, Méthodologie et applications [Revue] // Eyrolles. - 2002.

[29] Wu X, Zhu Y, A neural network regression model for relative dose computation [Revue] // Phys. Med.Biol.. - 2000. - Vol. 45 . - pp. 913–922 .

[30] Blake S.W. Artificial neural network modelling of megavoltage photon dose distributions [Revue] // Phys.Med. Biol. - 2004. - Vol. 49 . - pp. 2515–2526.

[31] Mathieu, Calculations of dose distributions using a neural network model [Revue] // Phys. Med. Biol. - 2005. - Vol. 50. - pp. 1019–1028.

[32] Bahi J, Neural network based algorithm for radiation dose evaluation in heterogeneous environments [Revue] // Int. Conf. on Artificial Neural Networks. - Athens, Greece : [s.n.], 2006. - Vol. 4132. - pp. 777–787.

[33] AAPM, Tissue inhomogeneity corrections for megavoltage photon [Rapport]. - 2004. - 85.

[34] Kalinin E, Deasy J: Amethod for fast 3-D IMRT dose calculations: The quadrant infinite beam (QIB) algorithm. Med Phys 2003, 30(6):1348–1349.

[35] Q. Hou, J. Wang, Y. Chen, and J.M. Galvin, “Beam orientation optimization for IMRT by a hybrid method of the genetic algorithm and the simulated dynamics,” Med. Phys. 30, 2360–2367 (2003).

[36] J. Stein, R. Mohan, X.H. Wang, T. Bortfeld, Q. Wu, K. Preiser, C.C. Ling, and W. Schlegel, "Number and orientations of beams in intensity-modulated radiation treatments," Med. Phys. 24, 149–160 (1997).

[37] C. Wang, J. Dai, and Y. Hu, "Optimization of beam orientations and beam weights for conformal radiotherapy using mixed integer programming," Phys. Med. Biol. 48(24), 4065–76 (2003).

[38] M. Ehrgott and R. Johnston, "Optimisation of beam directions in intensity modulated radiation therapy planning," OR Spectr. 25(2), 251–264 (2003).

[39] G.J. Lim, J. Choi, and R. Mohan, "Iterative solution methods for beam angle and fluence map optimization in intensity modulated radiation therapy planning," OR Spectr. 30(2), 289–309 (2007).

[40] S.X. Chang, T.J. Cullip, K.M. Deschesne, E.P. Miller, and J.G. Rosenman, "Compensators: an alternative IMRT delivery technique," J. Appl. Clin. Med. Phys. 5(3), 15–36 (2004).

[41] T.R. Mackie, T. Holmes, S. Swerdloff, P. Reckwerdt, J.O. Deasy, J. Yang, B. Paliwal, and T. Kinsella, "Tomotherapy: a new concept for the delivery of dynamic conformal radiotherapy," Med. Phys. 20(6), 1709–19 (1993).

[42] D.M. Shepard, M.A. Earl, X.A. Li, S. Naqvi, and C. Yu, "Direct aperture optimization: a turnkey solution for step-and-shoot IMRT," Med. Phys. 29, 1007–1018 (2002).

[43] M.A. Earl, D.M. Shepard, S. Naqvi, X.A. Li, and C.X. Yu, "Inverse planning for intensity-modulated arc therapy using direct aperture optimization," Phys. Med. Biol. 48, 1075–1089 (2003).

[44] M. Ehrgott, Ç. Güler, H.W. Hamacher, and L. Shao, "Mathematical optimization in intensity modulated radiation therapy," 4OR 6(3), 199–262 (2008).

[45] P. Kolmonen, J. Tervo, and T. Lahtinen, "Use of the Cimmino algorithm and continuous approximation for the dose deposition kernel in the inverse problem of radiation treatment planning," Phys. Med. Biol. 43(9), 2539–54 (1998).

[46] S. Webb, "Optimization of conformal radiotherapy dose distributions by simulated annealing: 2. Inclusion of scatter in the 2D technique," Phys. Med. Biol. 36(9), 1227–1237 (1991).

[47] T. Bortfeld, J. Bürkelbach, R. Boesecke, and W. Schlegel, "Methods of image reconstruction from projections applied to conformation radiotherapy," Phys. Med. Biol. 35(10), 1423–34 (1990).

[48] T. Bortfeld, “Optimized planning using physical objectives and constraints,” Semin. Radiat. Oncol. 9(1), 20–34 (1999).

[49] G.K. Bahr, J.G. Kereiakes, H. Horwitz, R. Finney, J. Galvin, and K. Goode, “The method of linear programming applied to radiation treatment planning,” Radiology 91(4), 686–93 (1968).

[50] D.M. Shepard, M.C. Ferris, G.H. Olivera, and T.R. Mackie, “Optimizing the Delivery of Radiation Therapy to Cancer Patients,” SIAM Rev. 41(4), 721–744 (1999).

Chapter II:

State of art of RADFET

Chapter II: State of art of RADFET

II.1. Introduction

Radiation therapy's primary objective is to increase the dosage to the tumor while reducing the exposure to normal surrounding tissues. Therefore, it is significant to be capable to quantify the radiation dose accurately and make sure the proper amount of radiation is delivered. The advancement of new radiation therapy technology, like intensity modulated radiotherapy [1] and image-guided radiation therapy [2] make dose verification a more and more complicated problem. The requirements of clinical radiation dosimetry for modern technology are as follows: the accuracy of the dose measurement has to be predictable up to 100 Gy, and the dose resolution should be as small as 1 cGy and position resolution within 1 mm. In this chapter, the emphasis will be on the RADFET dosimeter due to its later uses in this dissertation, and the other existing radiation dosimetry technology will be briefly summarized and their advantages and disadvantages will be discussed.

II.2. Radiation dosimetry technology

II.2.1.Optically stimulated luminescence dosimeter (OSLD)

Optically stimulated luminescence dosimeters (OSLDs) use the optically stimulated luminescence technique to detect radiation. The schematic representation of the energy levels of a crystalline material that maintains optical luminescence is shown in Figure II.1. Contaminations that create crystal-lattice defects are either present in pure crystalline dielectric materials or have been introduced in small amounts. These defects can function as electron or hole traps, as well as luminescence centers, emitting light when electrons and holes recombine around them. After radiation, free electrons and holes can be generated and trapped in the forbidden band (1-5 in Figure II.1). Figure II.1 encloses one-hole trap (2) as recombination center and three types of electron traps representing shallow traps (3), dosimetric traps (4) and deep traps (5). Shallow traps are unstable at room temperature and can only hold charge for very short periods of time. Deep traps can only release charge at very high temperature or stimulated with ultraviolet light. The dosimetric traps are energy-dense sufficient to keep the charge at ambient temperature for lengthy periods of time, but not so deep that the charge may be freed by visible light. Electron–hole recombination occurs after the trapped charges have escaped, resulting in light. The entire luminescence related to a specific level of traps is proportional to the trapped charge concentration and, in theory, to the absorbed radiation dosage [3–5].

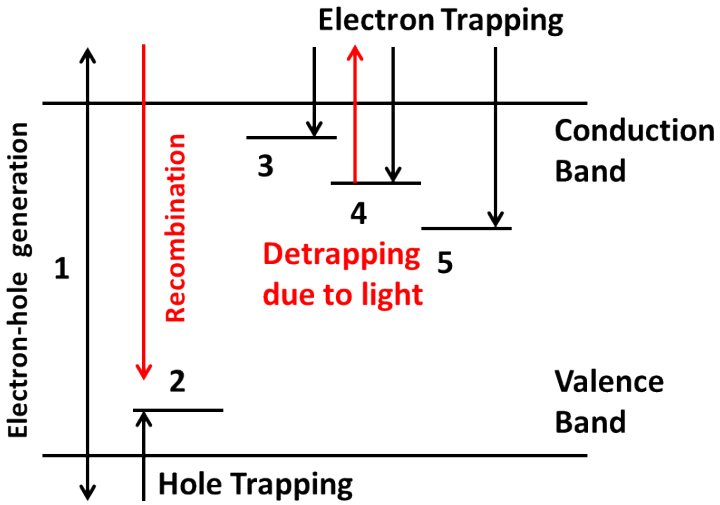


Fig.II.1. Schematic diagram of the energy levels of a crystalline material that sustains optical luminescence.

As a result, the trapped charge concentration in OSLDs serves as a record of the overall dosage absorbed by the crystal. This record can be "read" by stimulating the trapped charges using a light source to cause recombination of electron and hole and measuring the luminescence using a photomultiplier tube [6]. The advantage of OSLD is that it provides an accurate measurement for low radiation dose < 20 Gy with small or independent on beam quality, temperature, and angle of irradiation [7]. It also has a low radiation detectable threshold of 0.1 mGy[8] and is re-usable.

However, since OSLDs requires a photomultiplier tube to read the radiation dose, it cannot be used for real time measurements. They are also sensitive to light and therefore require extra packaging. Even though some groups have demonstrated a prototype OSLD dosimeter system for in vivo measurements [6,9], an optical fiber was needed to deliver the dosimeter into the body, making the measurement inconvenient and complicated.

II.2.2. Metal Oxide Semiconductor (MOS) capacitor

As an alternative to building the dosimeter using active electronic components, it is also possible to sense radiation passively. One approach is to use MOS capacitors as a radiation sensitive variable capacitor (varactor). Figure II.2 shows the structure and working mechanism of a MOS varactor. The varactor consists of a gate metal electrode, a thick SiO_2 layer for absorption of radiation and a lightly-doped n-type Si substrate. The total capacitance, C_{tot}measured from the gate and substrate is equal to the series combination of oxide capacitance, C_{ox} and silicon layer capacitance, C_{si}:

$$C_{tot} = \frac{C_{ox} \cdot C_{si}}{C_{ox}+C_{si}} \quad \text{(II.1)}$$

where C_{ox}, is a fixed value which depends on the oxide thickness, while C_{si} depends on the depletion layer thickness and can be modulated by gate bias or radiation. Radiation generated holes could be trapped at the SiO_2/Siinterface and decrease the depletion layer thickness in silicon. As a result, C_{tot} will change as a function of radiation. If connected with an inductor to form a resonant circuit, the resonant frequency will change as a function of radiation and can be potentially measured wirelessly.

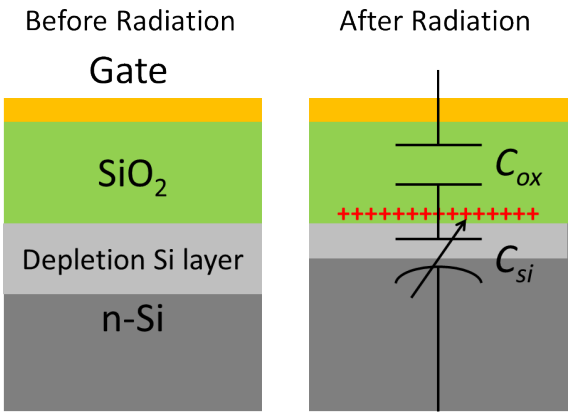


Fig.II.2. Structure of a MOS varactor and its working mechanism for radiation sensing.

Even though the MOS varactor enables the possibility of passive wireless sensing [10], [11], no wireless sensing has been demonstrated in the literature using a MOS varactor to date. One disadvantage of this structure is that it requires a thick SiO2 layer to be able to detect low doses of radiation. Since the capacitance is anti-proportional to SiO2 layer thickness, the capacitance per unit area of a MOS varactor is too low and this makes the device less scalable. Also, since the capacitance is modulated by the depletion capacitance, the tuning range of the total capacitance is small. This could limit the detection range of radiation.

II.2.3. Microelectromechanical(MEMS) technology

Another way to realize a passive varactor is to use MEMS technology. C. Son et al. demonstrated a microdosimeter composed of a radiation sensitive parallel plate capacitorand an inductor as shown in Figure II.3 [12]. In that work, the bottom plate of a capacitor is designed to be deflectable and an electret layer with pre-stored charge is placed under the top metal electrode. Radiation induced electron-hole pairs in the air gap will be collected from the electrode and reduce the surface charge density. As a result, the force between two parallel plates will decrease and the air gap will increase. In this way, the capacitance of the parallel plate capacitor will decrease as a function of radiation dose and the radiation can be detected as a changing of resonant frequency.

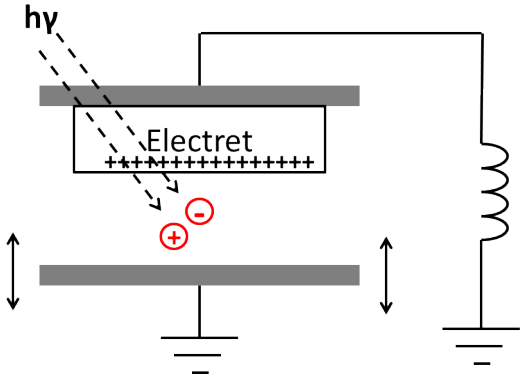


Fig.II.3. Structure of the MEMS microdosimeter

The dosimeter can be measured wirelessly and implanted with a hypodermic needle. Figure II.4 depicts the image of the dosimeter in a hypodermic needle (a) and a typical wireless measurement result with different capacitor sizes (b) [12]. The overall size of the dosimeter was 2.5 mm in diameter and 2.8 cm in length. Thus, is still too large to beimplanted easily and since the air gap between two plates is usually in micro-meters, the low capacitance per unit area makes further scaling difficult.

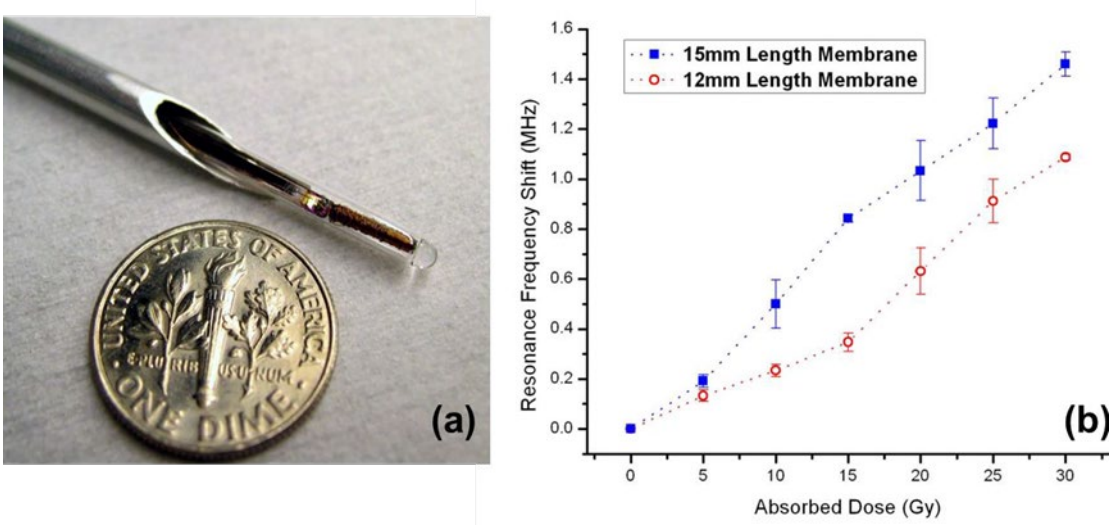


Fig.II.4. (a) Image of the MEMS dosimeter in a hypodermic needle. (b) typical wireless sensing results of the MEMS dosimeter with different capacitor size.

II.2.4. Radiation-Sensing Field-Effect-Transistor (RADFET)

Radiation-Sensing Field-Effect-Transistors In compared to diode dosimetry, RADFET detectors are relatively new for radiation treatment dosimetry. A typical RADFET is simply a p-MOSFET with a thick gate oxide as shown in Figure II.5. The oxide thickness varies from a few hundred nanometers to a few micrometers [13]. So, the charging of the MOSFET gate with accumulation charge created by ionizing radiation is the principle behind the operation of the MOSFET detector.

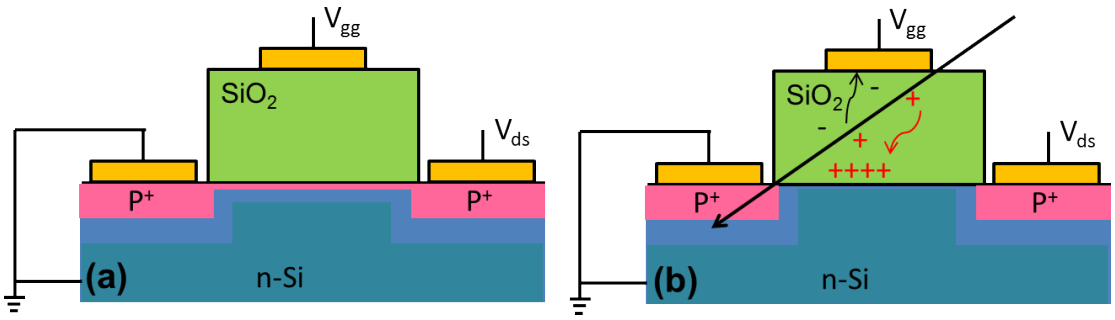


Fig.II.5. (a) The structure of a typical RADFET. (b) After irradiation, the trapped holes in SiO2 and Si interface will turn the device towards “off” state.

In fact, the use of MOSFETs for dosimetry was first suggested for tracking space radiation doses in order to forecast the integral dosage impact for satellite electronics [14].

- The RADFET detector has several benefits when used for dosimetry in radiation therapy:
- Extremely small size of dosimetric volume, which is impossible to have with other detectors;
- Like thermoluminescent dosimeters (TLDs), they can continuously retain the accumulated dosage and, unlike OSL detectors, can be readout without degradation of the dose information.
- They have a dosage rate independence of up to 108Gy/s.
- Because their sensitivity can be changed by gate bias, they are suited for a wide range of radiation applications.
- It may be readout after irradiation or in real time, allowing for real-time quality assurance or dose profiling in a water phantom [15].
- They are presently fairly inexpensive and can be disposable utilizing “one dose” principle.

All of the RADFET detector's advantages make them suitable for dosimetry in the domain of high electronic disequilibrium, such as on the body's surface or anatomical cavities, and in accumulation zones of depth dose curves in the case of irradiation on MV range X-rays on medical LINACs, which made us try to improve the RADFET performance, and propose a junctionless double graphene gate radiation sensitive FET (RADFET) in the fifth chapter of this dissertation.

The conversion of the threshold voltage shift ΔV_T into radiation dose D is used in ionizing radiation dosimetry employing radiation sensitive MOSFETs. The rise in the interface traps density and the positive trapped charge build-up or neutralization are caused by electron-hole pairs ofradiation-induced in the transistor's gate oxide layer (GOL).RADFETs' sensitivity may be modified, making them appropriate for a wide range of applications. Sensitivity can be adjusted, for example, by varying the thickness of the GOL [16,17], or, in some situations, by stacking transistors [14,15]. Positive bias on the gate during irradiation can also be used to modify the sensitivity [18,19].

II.3. Creation of defects precursors by ionizing radiation

Ionizing radiation causes a high number of defects at the interface of SiO_2 - Si and in SiO_2, which cause the threshold voltage change in MOSFETs. We'll go through the faults that have a big impact on the device's performance in more detail.

II.3.1. Ionization caused by photons

In SiO2 molecules, photons interact with the electrons during gamma or X-ray irradiation, producing SEs and holes, i.e. photons break Si_0 - $Si_0$and Si_0 - O covalent bonds in the oxide [20]. (The index o is used to indicate the presence of a silicon atom in an oxide). At the moment of creation, the highly energetic released electrons also known as "secondary electrons (SE)" may be recombined by holes, or they may evade recombination. SEs that manage to avoid recombination with holes go a considerable distancebefore leaving the oxide, wasting kinetic energy in collisions with bound electrons in the Si_0 - O and Si_0 - Si_0 covalent bonds, freeing more SEs (an oxygen vacancy is represented by the latter bond.).

Because the energy of a SE is habitually much higher than theenergy of an influence ionizing procedure, each SE can disruptmanyoxide covalent bonds before it exitor is recombined by the hole, producing severalnew secondary highly energetic electrons(For the formation of one electron-hole pair [20], i.e., ionization of the molecule, an energy of 18 eV is required).Because of the disparityin their effective cross sections, i.e., their effective masses, SEs are clearly more essential than highly energy photons in bond breakage. The electrons leaving the production site leave the oxide extremely quickly (a few picoseconds), but the holes stay.

Because there are no energetically deeper centers in the oxide bulk, the holes generated in the oxide bulk are typically only transitory, but not be stuck at the production sitepermanently. Depending on the direction of oxide electric field, the holes travel toone of the interfaces (SiO_2-Si or SiO_2-gate), where they are entrapped in energetically centers of deeper trap hole [21,22].Furthermore, even when the gate voltage is zero, the electrical potential created by a difference of work function amongthe gate and the substrate is sufficient to allow partial or total movement in the direction ofan interface.

II.3.2. The defects formedin impact ionization by SEs

By impact ionization, SEs travelling acrossthe oxide bulk disruptcovalent bonds and form the $\equiv Si_o - O^{\bullet+}Si_o \equiv$ complex. The symbol ≡ represents the three $Si_o - O$ bonds ($O_3 \equiv Si_o - O$) and $^\bullet$ signifies the electron that is not paired. The generated $\equiv Si_o - O^{\bullet+}Si_o \equiv$ combination, which represents the temporary hole center, is energetically very shallow (it is very easy for the trapped holes to escape. [23]).

The strained bondsilicon-oxygen $\equiv Si_o - O - Si_o \equiv$, which is mostly found near surfaces, can be easily disrupted by passing SEs, which commonly form non-bridging oxygen (NBO) centers, Si – O•, and positively charged E' centers, $\equiv Si_0^+$ [24], also known as E'_s centers [25]. An amphoteric defect called an NBO center can be more simplynegatively charged than positively charged when an electron is trapped. The NBO is the primary ancestorof traps (defects) withininterface regions and the oxide bulk due to its energetically deeper center.

A SE travelling across oxide can hit in the strained oxygen vacancy link $\equiv Si_o - Si_o \equiv$with an electron, which is anancestor to an E_γ' center ($\equiv Si_o^\bullet$), shattering the bond and hittingout an electron. Vacancy bonds of oxygen are primarily seen near interfaces.

The trapped charge may be positive (oxide trapped holes) or negative (oxide trapped electrons), with the firstbeing more significantbecause centers of hole trapping, such as E'_S, $E'_\gamma$and centers of NBO, are more numerous than electron trapping centers, with only one electron trap center (NBO). Because they have greatest impacton the channel carriers, the trapped electrons and holes closeto the Si – SiO_2 contact have the greatest impact on MOSFET properties.

II.3.3. Creation defects by hole transport in SiO_2

The trapped holes at$\equiv Si_o^+$ centers generated from oxygen vacancies and strained silicon-oxygen bonds are energetically profoundand stable, allowing them to remain unfilled for longer periods of time than superficiallytrapped holes. These centers can be found near both interfaces, particularly around the Si – SiO_2 one. Because there are many oxygen vacancies in addition tostrained silicon-oxygen bonds nearbyinterfaces, the holes formedand trapped at the bulk defects, representing energetically superficialcenters, are compelledto travel toone of the interfaces belowthe electric field, andthey are caughtat deeper traps. The holes abandonthe energetically shallow centers in the oxide spontaneously and travelto the interface as shown in figure.II.6 (a), via a jumpingprocess employing either superficialcenters in the oxide as

depicted in figure.II.6 (b); or centers in valence band of the oxide as illustrated in figure.II.6 (c) [21], [27]. Figure II.6 shows the in-spacehole transport for positive gate bias (a) and the diagram ofenergy for this space process various mechanisms (b), as well as the in-space hole transport for the negative gate bias(c).

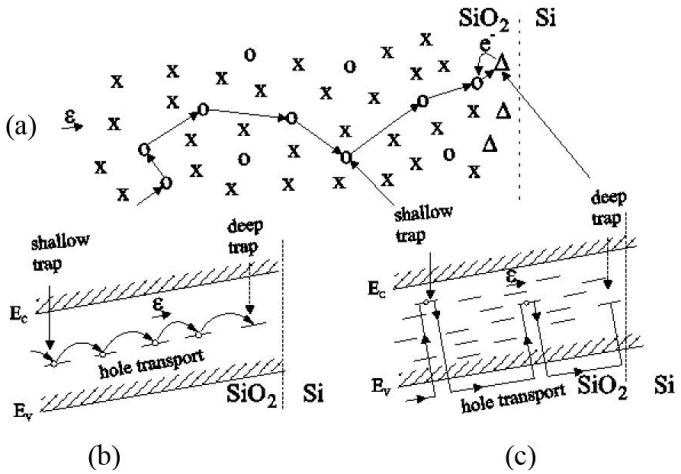


Fig.II.6.(a) Hole movement across the GOLwith positive gate bias. Unbroken and broken bonds (trapped holes at superficial traps) are represented by "x" and "o," respectively, and hole trap ancestors near the interface (ancestors of a deep trap)are represented by "Δ" (space diagram). (b) Hole movement by tunneling among localized traps and (c) by the oxide valence band (energetic diagram).

Figure II.7 illustrates the possibility of hole or electron to tunnel among two nearby centers: superficial and deep. It is impossible for holes or electrons to tunnel among these centers when there is no gate bias (Figure II.7 (a)). The bound electron can tunnel from the deep center to the surface center when the transistor is positively biased (Figure II.7 (b)). It symbolizes a hole tunneling from shallow to deep cores and being trapped at the deep center. In the surface center the electron, now, can easily tunnel from this surface center to the next modified surface center, allowing the hole to travel to the interface [21].

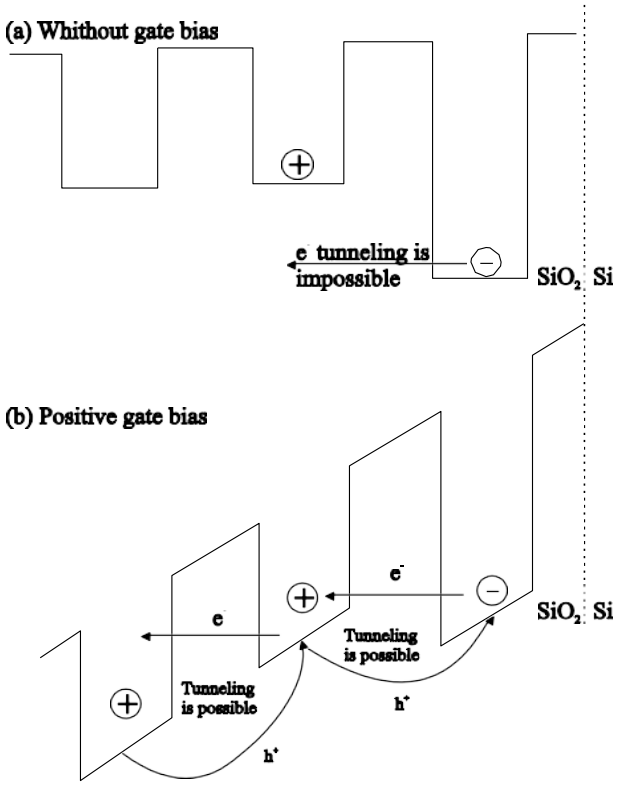


Fig.II.7. The electron tunneling among two nearby centers: (a) superficial and (b) deep.

The holes react with the hydrogen defects ≡ Si_o – H and ≡ Si_o – OH as they move through the oxide, eventually forming E'_S, E'_γ, NBO centers, hydrogen ions H^+, and hydrogen atoms H^o. At the SiO_2–Si contact, H^+ ions and H^o atoms were significant for defect formation (see the following section). When the holes reach the interface, they can break both the strained oxygen vacancy bonds ≡ Si_o – Si_o ≡ [24] and the strained silicon oxygen bonds ≡ Si_o – O – Si_o ≡ [21], generating E'_S and NBO centers. These centers, respectively, indicate energetically deeper hole and electron trapping centers. It should be emphasized that the energy levels of the defects formed after the holes at E'_S and E'_γ centers were trapped, as well as the electrons at the NBO center, can vary. Chemically identical flaws behave differently depending on the entire bond structure, including the angles and distances between the surrounding atoms [28-33].

II.3.4. Creation of SiO_2 - Si interface defects

True interface traps are amphoteric defects found at the SiO_2-Si interface. Si_3≡ $Si^\bullet$ (the index s denotes the silicon atom in the substrate): a silicon atom Si_3≡ $Si^\bullet$ back linked to three silicon atoms from the substrate ≡ Si_s commonly denoted as ≡ $Si^\bullet$ or $Si^\bullet$ at the SiO_2 -Si contact. They can be produced directly by incident photons passing through the substrate or gate [34,

35], although this amount can be ignored. Trapped holes (h^+ model) [36-39] and hydrogen released in the oxide (hydrogen-released species model– H model) [40-42] are the main sources of interface traps. According to the h^+ model, an interface trap was formed by a hole trapped near the SiO_2-Si interface, implying that an electron-hole recombination mechanism was involved [37]. When holes are trapped near the interface and electrons are supplied from the substrate, recombination occurs. The interface state can be created using the energy freed by the electron-hole recombination.

Under a positive electric field, H^+ ions generated in the oxide by trapped holes interact with with ≡ Si_o - H and ≡ Si_o - OH defects and drift toward the SiO_2-Si contact, according to the H model. The H^+ ion picks up an electron from the substrate at the contact, breaking a highly reactive hydrogen atom H^o [43]. The hydrogen atoms Ho produced in reaction holes with ≡ Si_o - H and≡Si_o - OH defects diffuse towards the SiO_2-Si interface under the existing concentration gradient, according to the H model. In interaction with interface trap precursors ≡ Si_s - H and ≡ Si_s - OH [44]-[46], these atoms react at the interface without an energy barrier, forming interface trap. Apart from the generation of interface traps in connection with interface trap precursors, interaction between H^o atoms with ≡ Si_s - H and≡ Si_s - OH precursors results in the formation of H_2 and H_2O molecules, respectively [21], [43]. H_2 molecules diffuse into the bulk of the oxide, cracking it at CC^+ centers [47]. This cracking process guaranteed a steady supply of H^+ ions, which drifted to the interface and formed interface traps [48].

II.3.5. Classification of defects based on their impact on I-V characteristics

Fixed traps (FT) and switching traps (ST) are two types of faults listed above (ST). FT denotes oxide traps that are unable to exchange charge with the channel (substrate) during the time period of the transfer/subthreshold characteristic measurement [49]. FTs can be negatively or positively charged, and the Coulomb force attracts or repels the channel carrier depending on the charge sign of both the FT and the channel carrier charge. ST denotes the traps formed near and at the SiO_2-Si interface, which collect (communicate with) the carrier from the channel during the transfer/subthreshold characteristic measurement time frame [49]. Slow switching traps (SST) are formed in the oxide near the SiO_2-Si interface, whereas fast switching traps (FST), also known as real interface traps, are formed at the interface. Slow states (SS) [50], anomalous positive charge (APC) [51,52], switching oxide traps (SOT) [53], and border traps [54] are all SSTs found in the oxide adjacent to the SiO_2-Si interface. The

influence of FT and ST on transistor subthreshold properties is shown as parallel shift and slope variation, respectively. FT are usually deeper in the oxide, and they can only be fully recovered or temporally compensated during the extended post-irradiation annealing process (as in the case of switching gate bias experiments). It is underlined that FST are amphoteric, and that each of them contributes to two states (an acceptor and a donor) within the silicon band gap, which might be randomly dispersed within it.

II.4. Characterization of transistor

There are numerous approaches for separating FT from ST [55]. Subthreshold midgap and charge pumping approaches are the most often employed techniques. Their fundamental concept will be presented.

II.4.1. Technique for a subthreshold midgap

The FT and ST densities are determined using the midgap-subthreshold (MG) approach [49], which is based on an investigation of MOSFET subthreshold properties. Specifically, the parallel shifts and slope changes of FT and ST on the transistor subthreshold characteristics in saturation effect the transistor subthreshold characteristics. The first phase, as depicted in Figure II.8, is linear regression of the linear regions of subthreshold features.

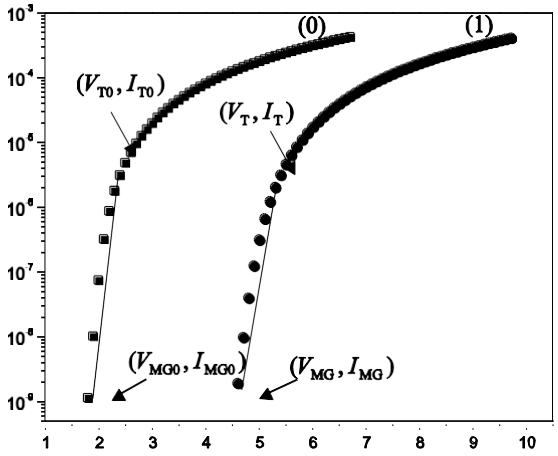


Fig.II.8. RADFETs Subthreshold properties with a 100 nm thick gate oxide made by Tyndall National Institute: (0) before irradiation with gamma-ray and (1) after 500 Gy irradiation.

A straight-line log *(ID)* = $M \times V_G + n$ is obtained via linear regression. The midgap current calculationI_{MG0} and I_{MG} before and after irradiation respectively, is the next step in the technique.

The subthreshold-current equation for a transistor in saturation [56] is used to calculate the midgap current:

$$I_D = \frac{\sqrt{2}\beta\varepsilon_S}{2C_{ox}L_D}\left(\frac{kTn_i}{qN_{A.D}}\right)^2\sqrt{\frac{kT}{q\psi_s}}e^{\left(\frac{q}{kT}\psi_s\right)} \qquad (.1)$$

Where $\beta = W\mu C_{ox}/L_{eff}$ and $L_D = \sqrt{\varepsilon_s kT}/\sqrt{q^2 N_{A.D}}$is the Debye length. In this equation W ,μ and C_{ox} denote the channel width, the carriers mobility and the oxide capacitance per unit area, respectively. L_{eff} denotes the effective channel length, ε_s is the silicon permittivity, k is the Boltzmann's constant, T refers to the absolute temperature, q, $N_{A,D}$ and n_i represent the electron charge, the doping concentration and the intrinsic carrier concentration. ψ_s denotes the surface potential.

When the surface potential s matches Fermi's potential F and Fermi's level is in the center of the semiconductor's energy gap, interface traps are electrically neutral (total charge equals zero) despite of the distribution across the substrate energy gap. The charge of FT alone produces a change towards the VG-axis of two subthreshold characteristics, and the gate voltage that corresponds to these surface potentials is denoted as VMG (midgap voltage) and may be computed as the abscissa of the (VMG, IMG) point at subthreshold characteristics (Figure II.8).

The V_{MG}, i.e., V_G that corresponds to $I_D = I_{MG}$ might be calculated as V_{MG} = [log (I_{MG}) - n]/ m using the equation log (I_D) = $m \times V_G + n$ derived by the linear fit of subthreshold characteristic. V_{MG0} and V_{MG} are discovered using this method. The straight lines acquired by the subthreshold characteristics linear fits are extended up to the matching midgap current I_{MG} in Figure II.8, which shows a region used for the linear fit. ΔV_{ft} is the component of threshold voltage shift caused by FT, it is expressed as:

$$\Delta V_{ft} = \Delta V_{MG} = V_{MG} - V_{MG0} \qquad (II.2)$$

where V_{MG0} and V_{MG} are the pre-irradiation and post-irradiation midgap voltages, respectively. The component of ST-induced threshold voltage shift, ΔV_{st}, is:

$$\Delta V_{st} = (V_T - V_{MG}) - (V_{T0} - V_{MG0}) = V_s - V_{s0} \qquad (II.3)$$

where V_{T0} and V_T are the threshold voltages of the transistors before and after irradiation, respectively, and threshold voltage shift is $\Delta V_T = V_T - V_{T0}$ .V_{T0}.

V_{T0} and V_T are calculated from the saturation transfer characteristics as the intersection of the V_G -axis with the extrapolated linear region of $\sqrt{I_D} = f(V_G)$ curves, which are described by the equation [56]:

$$I_D = \frac{\mu W C_{ox}}{2L_{eff}}(V_G - V_T)^2 \qquad \text{(II.4)}$$

The threshold voltage shift total value, ΔV_T is given by[57]:

$$\Delta V_T = \Delta V_{ft} + \Delta V_{st} \text{(II.5)}$$

$$\Delta V_T = \pm \frac{q}{C_{ox}} \Delta N_{ft} + \frac{q}{C_{ox}} \Delta N_{st} \text{(II.6)}$$

where $\Delta N_{ft}$and ΔN_{st} represents the FT areal density and theST areal density, respectively. P-channel and n-channel MOSFETs are denoted by the signs "+" and "-", respectively.

Expression (II.6) shows that the FT and ST both contribute in the same direction to the threshold voltage shift in p-channel MOSFETs. Furthermore, the "rebound effect" [20] doesn't really occur in p-channel MOSFETs; this effect is explained by competitive effects between the positive charge in the oxide and the negative interface traps produced in n-channel MOSFETs, leading in positive or negative VT values based on the different values of Nft and Nst. This is why p-channel MOSFETs are more typically utilized as ionizing radiation sensors or dosimeters. Because the carriers from the channel do not have enough time to reach them during measurement frames, ΔN_{ft} might contain a tiny quantity of SST that are situated deeper in the oxide.

II.4.2. Technique of charge pumping

Unlike the MG technique, the charge-pumping (CP) technique does not result in changes in charge densities in the positive oxide trapped charge and interface traps; instead, it is used solely to determine the density of interface traps, with the positive oxide trapped charge being determined later using the expression (II.6) if the change in threshold voltage is known [58-60].

Figure II.9 explains on the basis of the scheme the charge-pumping effect [59].

The transistor's source and drain are short-circuited, and the p-n junction of the source and drain with the substrate is polarized inversely with V_R voltage. When the signal is loss at the

gate, the inverted saturation current of these connections will flow due to inverted polarization at the junction source-substrate and drain-substrate. A shift in current direction in the substrate happens when a train of rectangular pulses of suitable amplitude is applied to the gate (using a pulse generator). The current intensity is proportional to the pulse frequency, and the same amount of electric charge is "pumped" towards the substrate. Because current cannot flow through oxide, the electric charge in the substrate passes through the source-drain p-n junction. In the case of n-channel MOSFETs, this results in the formation of a channel under the gate in the positive pulse half-period, where electrons are trapped on interface traps.

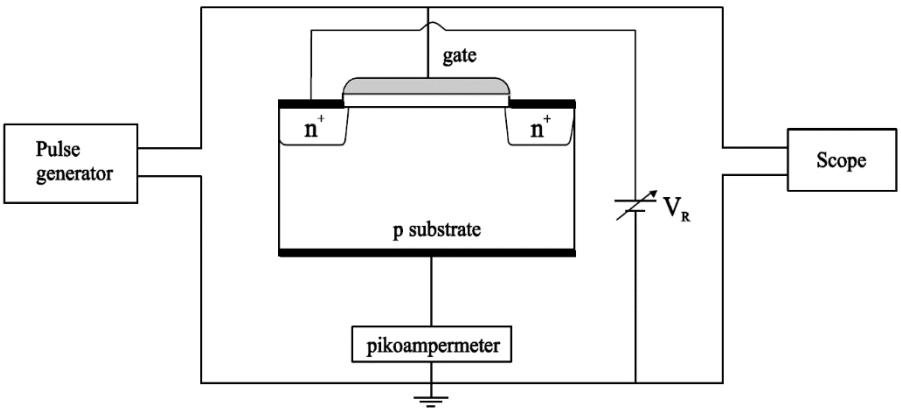


Fig.II.9. Charge pumping measurement shematic diagram.

When the channel area enters an accumulation state during the negative half-period, mobile electrons from the channel are returned to the source and drain, and the captured electrons are recombined with holes from the accumulated layer, resulting in the generation of CP current I_{CP}, whose maximum value $I_{CP,max}$ is expressed by[60]:

$$I_{CP,max} = fq^2 A_G \overline{D_{it}} \Delta\Psi_s = fqA_G \overline{D_{it}} \Delta E \qquad \text{(II.7)}$$

where f is the pulse frequency, A_G is the active charge pumping area under the gate and $\Delta\psi_S = q\Delta E$ is the surface potential complete sweep that matches to the ΔE. To avoid recombination with electrons of channel, ensure their return to the source and drain before cavities overflow from the substrate, which is achieved by using the p-n junction reverse polarization or a train of trapezoid or triangular pulses with sufficient rise t_r and fall t_f times pulse. However, a portion of the electrons whose capture is shallowest are thermally discharged into the substrate's conductive band, limiting the interface traps energy range width measured by the CP approach, resulting in CP current created by interface traps in the range of 0.5 eV from the forbidden band's middle. [60]

$$\Delta E = -2kTln(v_{th}n_i\sqrt{\sigma_n\sigma_p}\frac{|V_T-V_{FB}|}{|\Delta V_G|}\sqrt{t_r t_f}) \qquad \text{(II.8)}$$

In the expression (II.8), $\sigma_n$and $\sigma_p$are carrier captures cross section surfaces, v_{th} is thermal velocity, n_i is carrier self-concentration in the semiconductor, and ΔV_G is pulse height.

Equation (II.7) can be used to obtain the absolute value of interface traps density N_{it} and $N_{it} = \overline{D_{it}} \cdot \Delta E$:

$$N_{it} = \frac{I_{CPmax}}{q \cdot A_G \cdot f} \qquad \text{(II.9)}$$

The alteration in areal density of interface traps is$\Delta N_{it}(CP) = N_{it}(t) - N_{it0}$, where N_{it0} and $N_{it}(t)$ are the absolute value of interface trap density before irradiation and after irradiation time t respectively. Figure II.10 shows that I_{CPmax} is proportional to the pulse frequency, and a small-size transistor with normal state density requires at least several kHz for the charge-pump current level to reach the order of magnitude of pico-amperes.

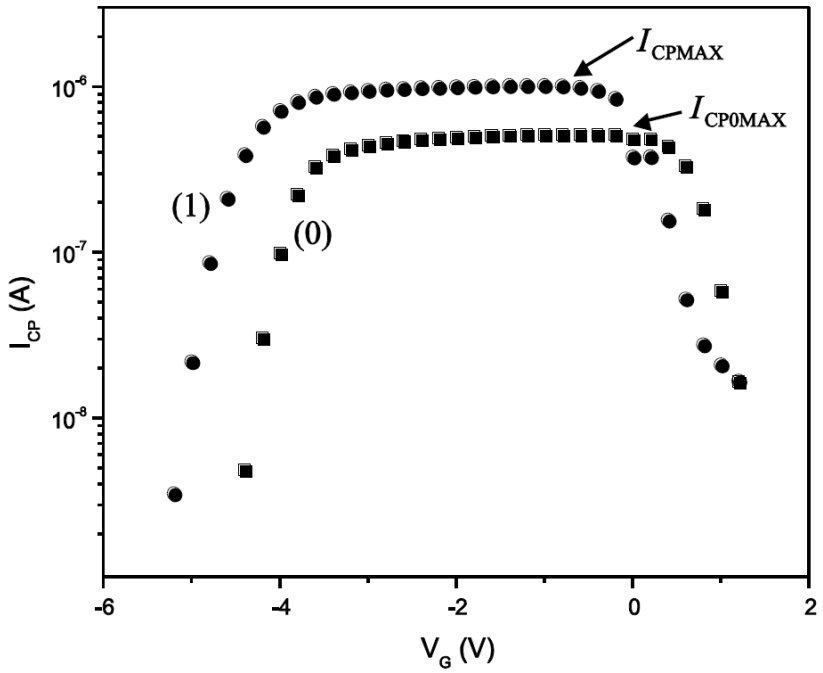


Fig.II.10. Tyndall National Institute RADFETs Elliot-tipe CP curves with a 100 nm thick gate oxide: (0) before gamma- ray irradiation and (1) after 500 Gyirradiation.

As a result, most CP measurements are done at frequencies between 100 kHz and 1 MHz, with only FST (true interface traps) being recorded (in some frequencies, CP is also contributed by of SST which also captures electrons from the channel [61]).Because the CP approach necessitated a separate substrate exit, it may be argued that it is incompatible with power VDMOSFETs in which the p-bulk is technologically coupled to the source. However, the CP approach for these devices can be used in a slightly different way [62-64].

II.4.3. measurements of threshold voltage shift at single point

One of the approaches for determining threshold voltage is based on transfer characteristics in saturation, which are defined as the intersection of the V_G-axis and the extrapolated linear area of,$\sqrt{I_D} = f\ (V_G)$ curves modeled by equation (II.4).

The drain-source voltage must be measured while the transistor is driven by a continuous drain current and the gate and drain terminals are short-circuited for the measurement of single point threshold voltage as shown in figure (II.11) [65]. The source-drain voltage shift is calculated as ΔV_T in this arrangement. During irradiation, the drain-source voltage can be continually monitored.

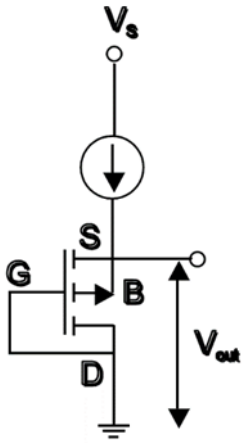
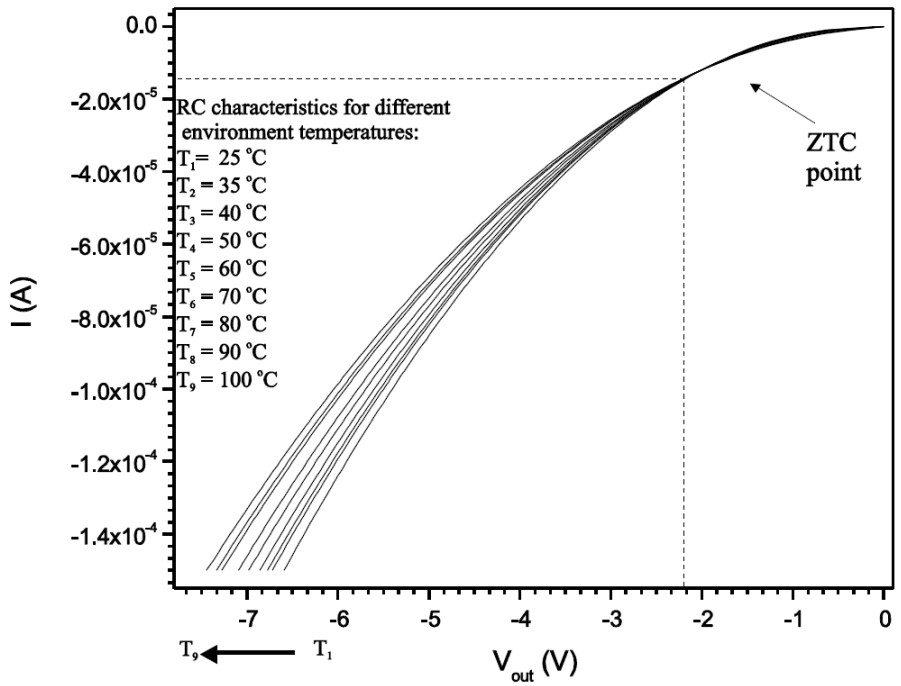


Fig.II.11. Configuration for measuring threshold voltage based on constant current.

Most commercial MOSFET-based dosimetry devices detect drain-source voltage increases at constant drain current [66-68]. Typically, the drain current chosen to minimize thermal drift is the zero-temperature coefficient current, I_{ZTC}, for which the drain-source voltage's thermal dependency cancels out. When I - V_{out} is measured at various temperatures, they all intersect at the same place. Figure II.12 shows readout currents ranging from 1 to 150

A and V_{out} voltage (V_{SD}) measurements for RADFETs with 400 nm thick gate oxide fabricated by Tyndall National Institute, at temperatures ranging from 25 to 100 °C. All of these curves converged in the neighborhood of 12 μA, as can be seen. It may be concluded that choosing



this current would reduce the temperature's effect on the threshold voltage.

Fig.II.12. RADFETs single-point characteristics at different temperatures with a 400 nm thick GOL made by Tyndall National Institute.

II.5. RADFET as ionizing radiation sensor and dosimeter

As previously indicated, the initial results in the use of MOSFETs in dosimetry were reported in 1974 [2]. The basic ideas for using these devices as ionizing radiation sensors and dosimeters were discussed. Following that, a number of research groups dealing with comparable issues arose. Canadian [69], United States Navy [70,71], French [72,73], Netherlands [74,75], United States [76,77] and Serbia [78-80] are among them.

Radiation sensitive MOSFETs are manufactured by a large number of companies and institutes around the world. In Cork, Ireland, Tyndall National Institute, is one of them. RADFETs with gate oxide thicknesses of 100 nm, 400 nm, and 1 μm are manufactured at this facility. This presentation will show some of the results relating to these components, as well as numerous critical dosimetric characteristics.

II.5.1. RADFETs re-use possibility

Many studies have shown that RADFETs cannot be used to determine the dose of ionizing radiation afterward. Specifically, these dosimeters are solely used to estimate the maximum dose, which is dictated by the RADFET type and sensitivity. These RADFETs should be replaced once the maximum radiation dose has been attained. The initial results on the possibility of reusing these devices are presented in [10] for a radiation exposure of 400 Gy. [13], [81] present follow-up experiments for the same components. Irradiation was done with gamma rays up to 35 Gy, with and without gate bias (V_{irr}= 2.5V and V_{irr}= 5V). Figure II.13 depicts the threshold voltage shift ΔV_T as a function of radiation dose D for both the first and second irradiation with V_{irr}= 5V gate bias. The RADFETs were annealed at room temperature for 5232 hours without gate bias after the first irradiation. The annealing process was then conducted for 432 hours at 120 °C without gate bias. The RADFETs were then exposed to the same radiation. The ΔV_T levels during the first and second irradiation are extremely similar. These findings contradict previous findings [10], which showed that ΔV_T values acquired during the first irradiation are higher than those obtained after the second irradiation.

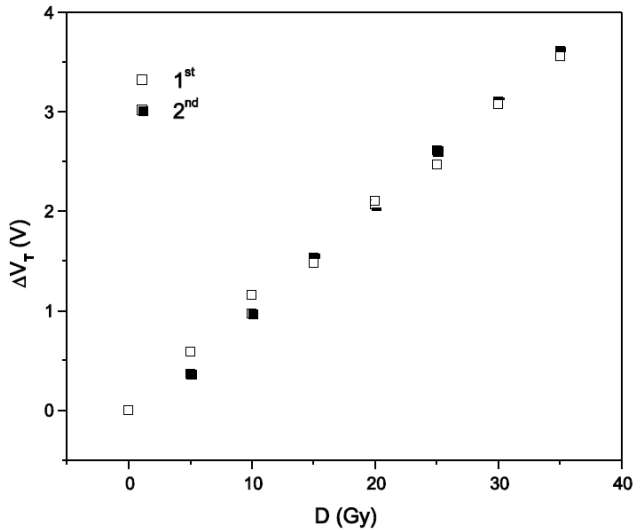


Fig.II.13. Threshold voltage changeΔV_Tas a function of radiation dose D of 400 nm thick GOLRADFETs with V_{irr}= 5V gate bias for both the first irradiation and second irradiation.

The increase in ΔN_{ft}as depicted in figure II.14 is nearly identical after the first and second irradiation of RADFETs, but the increase in ΔN_{st} (MG) is larger during the second irradiation as shown in figure II.15. During the second irradiation, ΔN_{fst} (CP) is higher as illustrated in figure II.16. Figures II.14, II.15, and II.16 show that at a radiation dosage of 35 Gy, the predominant contribution to ΔV_T rise during the first and second irradiation comes from FT, which has a

density that is an order of magnitude larger than ST (MG) density.

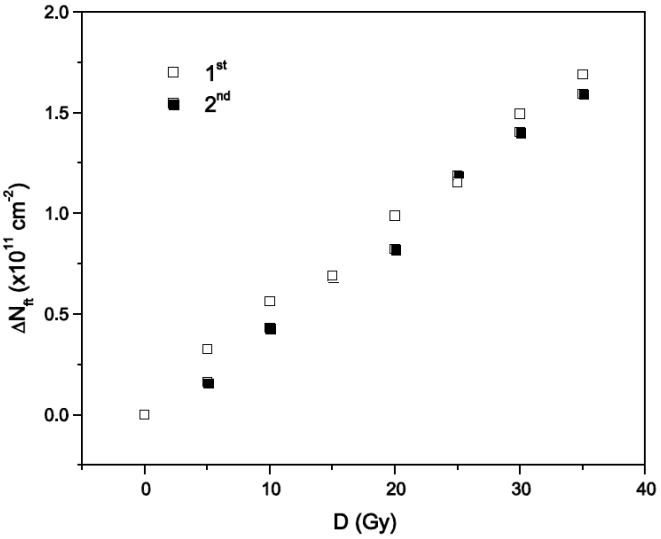


Fig.II.14. Fixed traps ΔN_{ft}Areal density as a function of radiation dose D of 400 nm thick GOLRADFETs with V_{irr}= 5V gate bias for both the first and second irradiation.

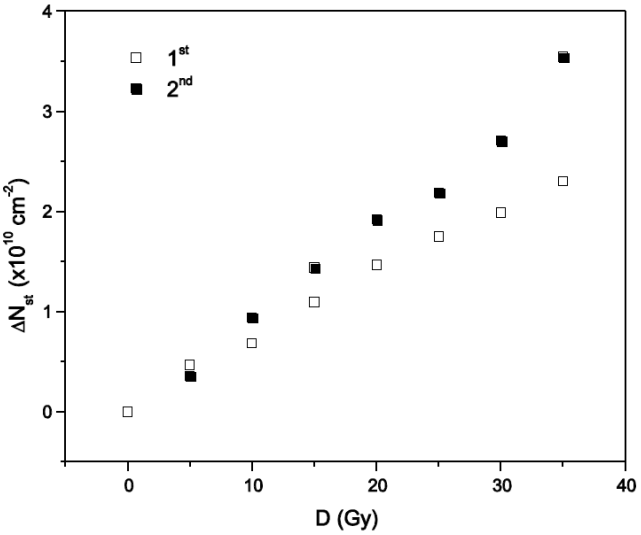


Fig.II.15. Switching traps ΔN_{st}(MG)areal density as a function of radiation dose D of 400 nm thick GOLRADFETs with V_{irr}=5V gate bias for both the first and second irradiation.

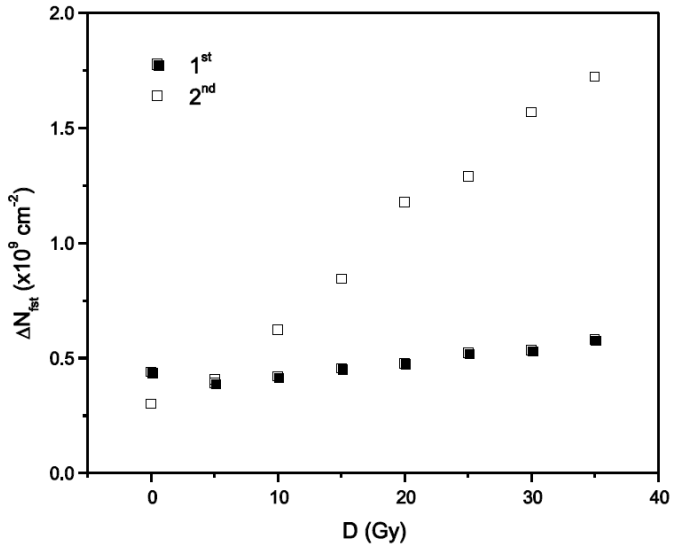


Fig.II.16. Switching traps ΔN_{st}(CP) areal density as a function of radiation dose $D$of 400 nm thick GOLRADFETs with V_{irr}=5V gate bias for both the first and second irradiation.

II.7. Conclusion

In order to examine the applicability of radiation sensitive MOSFETs (RADFETs) in dosimetry, much research has been conducted. Their tiny volume gives them a benefit over other dosimetric systems, which is especially relevant in in-vivo dosimetry and the management of x-ray gradient radiation fields. The most prevalent applications for RADFETs are photon and ionizing radiation charged particle detection. It can also be used to detect neutrons, however their sensitivity is far lower than that of photons or charged particles. Their sensitivity can be improved by applying gate bias during irradiation and thickening the GOL. The sensitivity increases as the photon energy of ionizing radiation decreases. These components must accomplish little variation in threshold voltage shift after irradiation at room temperature, i.e., the dosimetric information must be preserved for a long period of time. Because they can register dosages as low as 1 cGy, RADFETs are considered sensitive gamma and x-ray sensors. Unfortunately, one of their significant drawbacks is rapid fading following irradiation. Some commercially available p-channel MOSFETs have been demonstrated to be very effective as gamma and x-ray sensors, as well as electron sensors with energy of several MeV, in recent studies. 3N163, DMOS BS250F, ZVP3306, ZVP4525, and power VDMOSFETs IRF9520 are low power p-channel MOSFETs. In addition, p-channel MOS transistors, such as the CD4007, can be employed as ionizing radiation sensors.

Radiation detectors, which are an important aspect of radiation therapy, are required for dose verification to ensure that the delivery of radiation to the target is proceeding as planned. Due to its tiny size, ability to read data in real time, separation of components of mixed radiation fields, biological dosimetry on a cellular level, and dose imaging on medical LINACs, RADFETs dosimeters have several advantages in this application.

Referencs

[1] Y. Nishimura and R. Komaki, Intensity-modulated radiation therapy: clinical evidence and techniques. Springer, 2015.

[2] J. D. Bourland, Image-guided radiation therapy Imaging in medical diagnosis and therapy. CRC Press, 2012.

[3] J. M. Edmund and C. E. Andersen, "Temperature dependence of the Al2O3:C response in medical iuminescence dosimetry," Radiat. Meas., vol. 42, no. 2, pp. 177–189, 2007.

[4] B. G. Markey, S. W. S. McKeever, M. S. Akselrod, L. Botter-Jensen, N. Agersnap Larsen, and L. .Colyyott, "The temperature dependence of optically stimulated luminescence from Alpha-Al2O3:C," Radiat. Prot. Dosimetry, vol. 65, pp. 185–189, 1996.

[5] E. G. Yukihara, V. H. Whitley, J. C. Polf, D. M. Klein, S. W. S. McKeever, A. E. Akselrod, and M. S. Akselrod, "The effects of deep trap population on the thermoluminescence of Al 2O3:C," Radiat. Meas., vol. 37, no. 6, pp. 627–638, 2003.

[6] E. G. Yukihara and S. W. S. McKeever, "Optically stimulated luminescence (OSL) dosimetry in medicine.," Phys. Med. Biol., vol. 53, pp. R351–R379, 2008.

[7] P. A. Jursinic, "Characterization of optically stimulated luminescent dosimeters, OSLDs, for clinical dosimetric measurements.," Med. Phys., 2007.

[8] C. A. Perks, C. Yahnke, and M. Million, "Medical dosimetry using Optically Stimulated Luminescence dots and microStar readers," in 12th International Contress of the International Radiation Protection Association, 2008.

[9] C. E. Andersen, S. K. Nielsen, S. Greilich, J. Helt-Hansen, J. C. Lindegaard, and K. Tanderup, "Characterization of a fiber-coupled Al2O3:C luminescence dosimetry system for online in vivo dose verification during 192Ir brachytherapy.," Med. Phys.

[10] "US20100219494A1," 2010.

[11] M. Gopalan, "Experimental study of MOS capacitors as wireless radiation dose sensors," 2010.

[12] C. Son and B. Ziaie, "A wireless implantable passive microdosimeter for radiation oncology.," IEEE Trans. Biomed. Eng., vol. 55, no. 6, pp. 1772–1775, Jun. 2008.

[13] A. Holmes-Siedle and L. Adams, "RADFET: A review of the use of metal-oxide-silicon devices as integrating dosimeters," Int. J. Radiat. Appl. Instrumentation. Part C. Radiat. Phys. Chem., vol. 28, no. 2, pp. 235–244, 1986.

[14] Holmes-Siedle, "The Space Charge Dosimeter", Nucl. Inst. Meth, 121, 169–179, (1974).

[15] Mandal A, Ram C, Mourya A, Singh N. Small field depth dose profile of 6 MV photon beam in a simple air-water heterogeneity combination: A comparison between anisotropic analytical algorithm dose estimation with thermoluminescent dosimeter dose measurement. J Cancer Res Ther. 2017.

[16] G. Ristić, S. Golubović and M. Pejović, „pMOS dosimeter with two-layer gate oxide operated at zero negative bias", Electr. Lett., vol. 30, pp. 295-296, 1994.

[17] G. Ristić, A. Jakšić, M. Pejović, "pMOSdosimetric transistors with two-layer gate oxide", Sensors and Actuators A, vol. 63, pp. 129-134, 1997.

[18] G. Sarrabayrouse and F. Gessinn, "Thick oxide MOS trnsistors for ionizing radiation dose measurement", radioprotection, vol. 29, pp. 557-572, 1994.

[19] A. Haran, A. Jakšić, N. Rafaeli, A. Elyahu, D. David and J. Barak, IEEE Trans. Nucl. Sci., vol. 51, 2917-2921, 2004.

[20] T. P. Ma and P.V. Dressendorfer, Ionizing Radiation Effects in MOS Devices and Circuits, New York: Willey and Sons, 1989.

[21] G. S. Ristić, "Influence of ionizing radiation and hot carrier injection on metal-oxide-semiconductor transistors", J. Phys. D: Appl. Phys., vol. 41, 023001 (19 pp), 2008.

[22] M. Pejović, P. Osmokrović, M. Pejović and K. Stanković, "Influence of ionizing radiation and hot carrier injection on metal-oxide-semiconductor transistors". In M. Nenoi (Ed), Current Topic in Radiation Research. INTECH. Institute for New Technologies, Maastricht (NL), Chapter 33, <http://www.intechopen. com/books/current-topics-in-ionizing-radiation-research>. OCLC: 846871029, 2012.

[23] C. T. Sah, "Origin of interface states and oxide charges generated by ionizing radiation", IEEE Tran. Nucl. Sci., vol. 23, pp. 1563-1567, 1976.

[24] D. L. Griscom, "Optical properties and structure of defects in silica glass", J. Ceram. Soc. Japan, vol. 99, pp. 923-941, 1991.

[25] R. Helms and E.H. Poindexter, "The silicon-silicon-dioxide system: its microstructure and imperfections", Rep. Prog. Phys., vol. 57, pp. 791-852, 1994.

[26] R. A. Weeks, "Paramagnetic resonance of lattice defects in irradiated quartz", J. Appl. Phys., vol. 27, pp. 1376-1381, 1959.

[27] H. E. Boesch, Jr, F.B. McLean, J.M. McGarrity and G.A. Ausman, Jr,"Hole transport and charge relaxation in irradiated SiO2 MOS capacitors", IEEE Trans. Nucl. Sci., vol. 22, pp. 2163-2167, 1975.

[28] W. L. Warren and P.M. Lenahan, "A comparison of positive charge generation in high field stressing and ionizing radiation on MOS structure", IEEE Trans. Nucl. Sci., vol. 34, pp. 1355-1358, 1987.

[29] L. P. Trombetta, F.J. Feigl and R.J. Zeto, "Positive charge generation in metal-oxide-semiconductor capacitors, J. Appl. Phys., vol. 69, pp. 2512-2521, 1991.

[30] R. K. Freitag, D.B. Brown and C.M. Dosier, "Experimental evidence of two species of radiation induced trapped positive charge", IEEE Trans. Nucl. Sci., vol. 40, pp. 1316-1322, 1993.

[31] R. K. Freitag, D.B. Brown and C.M. Doser, "Evidence for two types of radiation-induced trapped positive charge", IEEE Trans. Nucl. Sci., vol. 41, pp. 1828-1834, 1994.

[32] J. E. Conley, P.M. Lenahan, A.H. Lelis and T.R. Oldham, "Electron spin resonance evidence for the structure of a switching oxide trap: long term structural charge at silicon dangling bond sites in SiO2", Appl. Phys. Lett., vol. 67, pp. 2179-2181, 1995.

[33] J. F. Conley, P.M. Lenahan, A.J. Lelis and T.R. Oldham, "Electron spin resonance evidence that center can behave as switching oxide trap", IEEE Trans. Nucl. Sci., vol. 42, pp. 1744-1749, 1995.

[34] D. A. Buchanan, A.D. Marwick, D.J. DiMaria and L. Dori, "Hot-electron-induced hydrogenredistribution and defect generation in metal-oxide-semiconductors", J. Appl. Phys., vol. 76, pp. 3595- 3605, 1994.

[35] D. J. DiMaria, D.A. Buchanan, J.H. Stathis and R.E. Stahlbush, "Interface states induced by the presence of trapped holes near the silicon-silicon-dioxide interface", J. Appl. Phys., vol. 77, pp. 2032- 2040, 1995.

[36] S.K. Lai, "Two carrier nature of interface-state generation in hole trapping and radiation damage", Appl. Phys. Lett., vol. 39, pp. 58-60, 1981.

[37] S. K. Lai, Interface trap generation in silicon dioxide when electrons are captured by trapped holes", J. Appl. Phys., vol. 54, pp. 2540-2546, 1983.

[38] S. T. Chang, J.K. Wu and S.A. Lyon, "Amphoterical defects at Si-SiO2", Appl. Phys. Lett., vol. 52, pp. 622-624, 1986.

[39] S. J. Wang, J.M. Sung and S.A. Lyon, "Relationship between hole trapping and interface state generation in metal-oxide-silicon structures, Appl. Phys. Lett., vol. 52, pp. 1431-1433, 1986.

[40] F. B. McLean, "A framework for understanding radiation-induced interface states in SiO2 MOS structures, IEEE Trans. Nucl. Sci., vol. 27, pp. 1651-1657, 1980.

[41] N. S. Saks, C.M. Dozier and D.B. Brown, "Time dependence of interface trap formation in MOSFETs following pulsed irradiation", IEEE Trans. Nucl. Sci., vol. 35, no. 6, pp. 1168-1177, 1988.

[42] N. S. Saks and D.B. Brown, "Interface trap formation via the two-stage H+ process", IEEE Tran. Nucl. Sci., vol. 36, no. 6, pp. 1848-1857, 1989.

[43] D. L. Griscom, D.B. Brown and N.S. Saks, Nature of radiation-induced point deffcts in amorphous SiO2 and their role in SiO2-on-Si structure,The Physics and Chemistry of SiO2 and Si- SiO2 interface, ed C.R. Holmes and B.E. Deal, Ney-York, Plenum, 1988.

[44] K. L. Brower and S.M. Mayers, "Chemsical kinetics of hydrogen and (111) Si- SiO2 interface defect", Appl. Phys. Lett., vol. 57, pp. 162-164, 1990.

[45] J. H. Stathis and E. Cartier, "Atomic hydrogen reactions with Pb centers at the (100) Si- SiO2 interface", Phys. Rev. Lett., vol. 72, pp. 2745-2748, 1994.

[46] E. H. Poindexter, "Chemical reactions of hydrogenous species in the Si- SiO2 system", J. Non. Cryst. Solids, vol. 187, pp. 257-263, 1995.

[47] R. E. Stahlbush, A.H. Edwards, D.L. Griscom and B.J. Mrstik, "Post-irradiation cracking of H2 and formation of interface states in irradiated metal-oxide-semiconductor field-effect transistors", J. Appl. Phys., vol. 73, pp. 658-667, 1993.

[48] M. M. Pejović, "Physico-chemical processes in vertical-double-diffusion metal-oxide-semiconductor field effect transistors induced by gamma-ray irradiation and post-irradiation annealing", Facta Universitatis, Series: Physics, Chemistry and Technology, vol. 13, pp. 13-27, 2015.

[49] McWhorter and P.S. Winocur, "Simple technique for separating the effects of interface traps and trapped- oxide charge in metal-oxide semiconductor transistors", Appl. Phys. Lett., vol. 48, pp. 133-135, 1986.

[50] M. V. Fischetti, R. Gastaldi, F. Maggoni and A. Madelli, "Slow and fasdt states induced by hot electrons at Si- SiO2 inteface", J. Appl. Phys., vol. 53, pp. 3136-3144, 1982.

[51] L. P. Trombetta, F.J. Feigl and R.J. Zeto, "Positive charge generation in metal-oxide-semiconductor capacitors", J. Appl. Phys., vol. 69, pp. 2512-2521, 1991.

[52] R. K. Freitag, D.B. Brown and C.M. Dozier, "Experimental evidence of two species of radiation induced trapped positive charge", IEEE Tran. Nucl. Sci., vol. 40, pp. 1316-1322, 1993.

[53] A.J. Lelis. and T.R. Oldham, "Time dependence of switching oxide traps", IEEE Tran. Nucl. Sci., vol. 41, pp. 1835-1843, 1994.

[54] D. M. Fleetwood, "Border traps in MOS devices", IEEE Tran. Nucl. Sci., vol. 39, 269-271, 1992.

[55] V. Davidovic, Ph. D., University of Nis, 2010.

[56] S. M. Sze, Physics of Semiconductor Devices, Ney York, Wiley, 1981.

[57] A. Holmes-Siedle and L. Adams, Handbook of Radiation Effects, 2nd ed., New York: Oxford University Press, 2002.

[58] M.A.B. Eliot, "The use charge pumping currents to measure surface state densities in MOS transistors",Solid-State Electron., vol. 19, pp. 241-247, 1986.

[59] J.S. Brugler and P.G. Jespres, "Charge pumping in MOS devices", IEEE Trans. Electron DEv. Lett., vol. 13, pp. 627-629, 1969.

[60] G. Groeseneken, H.E. Maes, N. Baltron and R.F. De Keersmaeeker, "A reliable approch to charge- pumping measurements in MOS transistors", IEEE Trans. Electron Dev., vol. 31, pp. 42-53, 1984.

[61] R. E. Paulsen, R.R. Siergiej, M.L. French andM.H. White, "Observation of near-interface oxide traps with the change pumping technique", IEEE Electron Dev. Lett., vol. 13, pp. 627-629, 1992.

[62] D. Habaš, Z. Prijić, D. Pantić and N. Stojadinović, "Charge-pumping characterization of SiO2/Si interface virgin and irradiated power VDMOSFETs", IEEE Trans. Electron Dev., vol. 43, pp. 2197-2208, 1996.

[63] S. C. Witezak, K.F. Gallawoy, R.D. Schrimpf and J.R. Brews, G. Prevost, " The determination of Si- SiO2 interface trap density in irradiated four-terminal VDMOSFETs using charge pumping", IEEE Trans. Nucl. Sci., vol. 43, pp. 2558-2564, 1996.

[64] G. S. Ristić, M.M. Pejović and A.B. Jakšić, „Comparison between post-irradiation annealing and post- high electrical field stress annealing of n-channel power VDMOSFETs", Appl. Surf. Sci., vol. 220, pp. 181-185, 2003.

[65] A. Kelleher, M. O'Sullivan, J. Rayn, B. O'Neal and W. Lane, "Development of the radiation sensitivity of pMOS dosimeters", IEEE Tran. Nucl. Sci., vol. 39, pp. 342-346, 1992.

[66] I. Thomson, "Direct reading dosimeters", European Patent Office, Ep0471957A2, 02/07/1991.

[67] S. Best, A. Ralson and N. Suchowerska, “Clinical application of the one dose patient dosimetry system for total body irradistion”, Phys. in Medic. and Biology, vol. 50, pp. 5909-5919, 2005.

[68] M. M. Pejović, “The gamma-ray irradiation sensitivity and dosimetric information instability of RADFET dosimeter”, Nucl. Technol. and Radiat. Protection, vol. 28, pp. 415-421, 2013.

[69] I. Thomson, R.E. Thomson and L. P. Brendt, “Radiation dosimetry with MOS sensors”, Radiation Protec. Dosimetry, vol. 6, pp. 121-124, 1983.

[70] L. S. August, R.R. Circle and J.C. Ritter, “An MOS dosimeter for use in space”, IEEE Tran. Nucl. Sci., vol. 30, pp. 508-511, 1983.

[71] L. S. August, “Estimating and reducing errors in MOS dosimeters caused by exposure to different radiations”, IEEE Trans. Nucl. Sci., vol. 29, no. 6, pp. 2000-2003, 1982.

[72] G. Sarrabayrouse, A. Bellaouar and P. Rossel, “Electrical properties of MOS radiation dosimeters”,Revue Phys. Appl., vol. 21, pp. 283-287, 1986.

[73] A. Ballaouar, G. Sarrabayrouse and P. Rassel, “MOS transistor for ionizing radiation dosimetry”, Proc. 13th Yugoslav Conf. on Mictoelectronics (MIEL 85), Ljubljana, pp. 161-168, 1985.

[74] L. Adams and A. Holmes-Siedle, “The development of MOS dosimetry unit for use in space”, IEEE Trans. Nucl. Sci., vol. 18, pp. 1607-1612, 1978.

[75] L. Adams, E.J. Daly, R. Harboe-Sorensen, A.G. Holmes-Siedle, A.K. Ward and A.A. Bull, “Measurements of SEU and total dose in geostationary orbit under normal and solar frame conditions”, IEEE Trans. Nucl. Sci., vol. 38, pp. 1686-1692, 1991.

[76] J. S. Leffler, S.R. Lendgren and A.G. Holmes-Siedle, “The aplications of RADFET dosimetry to equipment radiation qualification and monitoring”, Trans. of the American Society, vol. 60, pp. 535- 536, 1989.

[77] A. G. Holmes-Siedle, L. Adams, J.S. Leffler and S.R. Lingren, ”The RADFET system for real-time dosimetry in nuclear facilities”, 7th Annual ASTM-Euratom Symp. on Reac. Dosimetry, Strasbourg, pp. 851-859, 1990.

[78] G. Ristić, S. Golubović and M. Pejović, "P-channel metal-oxide-semiconductor detector fading dependencies on gate bias and oxide thickness", Appl. Phys. Lett., vol. 66, pp. 88-89, 1995.

[79] G. Ristić, S. Golubović and M. Pejović, "Sensitivity and fading of pMOS dosimeters with thick gate oxide", Sensors and Actuators A, vol. 51, pp. 153-158, 1996.

[80] Z. Savić, S. Stanković, M. Kovačević and M. Petrović, „Energy dependence of pMOSdosimeters", Radiation Protect. Dosimetry, vol. 64, pp. 205-211, 1996.

[81] M. M. Pejović, M. M. Pejović and A.B. Jakšić, "Response of pMOS dosimeters on gamma-ray irradiation during its re-use", Radiation Protection Dosimetry, vol. 155, pp. 394-403, 2013.

Chapter III:

Multiobjective evolutionary algorithms

Chapter III: Multiobjective evolutionary algorithms

III.1. Introduction

Multi-objective optimization is an important element of optimization activities because virtually all real-world optimization issues are perfectly appropriate to being represented by means of several competing objectives. The traditional approach to addressing such issues was largely centered on scalarizing many objectives into a single target, but the evolutionary approach has been to tackle a Multi-objective Optimization (MO) problem as is. This chapter discusses the underlying concepts of MO, the distinctions of multi-objective optimization and single-objective optimization, and some well-known standard and evolutionary MO methods. Two case examples demonstrate the value of MO in reality. Following that, a variety of research problems are mentioned. The chapter closes by identifying certain significant resources funds to the subject of MO and recommending some few tips and tricks.

Most real-world search and optimization issues are naturally framed as non-linear programming problems with many competing goals. Caused by an absence of appropriate solution methodologies, such issues were intentionally transformed into and addressed as a single-objective problem. The problem arises since such situations provide a group of trade-off optimal solutions (called Pareto-optimal solutions), rather than a single optimum solution. It is thus critical to identify plenty of Pareto-optimal solutions as feasible, rather than just one. This fact due to that any two such solutions represent a trade-off between the objectives, and when such trade-off solutions are revealed, consumers will be in a better position to make a decision.

Conventional techniques approach similar issues with a different mindset, owing to a lack of an appropriate optimization tool for effectively finding numerous optimal solutions. They often need several uses of an algorithm to identify numerous Pareto-optimal solutions, and such applications often do not ensure the discovery of every Pareto-optimal answers. In contrast, the population method of evolutionary algorithms (EAs), on the other hand, is a fast technique to identify many Pareto-optimal solutions in a single simulation run. This characteristic has rendered evolutionary multi-objective optimization (EMO) research and uses prominent during the last 15 years. The inquisitive reader can investigate current research topics and other significant research in a variety of books [13], conference proceedings [4, 5], and countless research articles (archived and maintained in [6].

In this lesson, we will look at the key distinctions between single-objective and multi-objective optimization problems. The optimality requirements in a multi-objective optimization problem are explained, and a variety of cutting-edge multi-objective optimization approaches, including one evolutionary method, is provided. We offer a handful of fascinating research papers to illustrate that evolutionary multi-objective techniques are capable and suited for tackling real-world issues. Lastly, a variety of different EMO research issues are highlighted.

A multi-objective optimization problem (MOP) considers several objective functions. Numerous objectives or multiple criteria are present in the majority of actual decision-making issues. In the previous, a MOP was generally cast and addressed as a single-objective optimization problem due to a lack of acceptable solution techniques. Nevertheless, there are a plethora of different discrepancies in the operating mechanism of single- and multi-objective optimization algorithms, which necessitates the employment of a multi-objective optimization approach to solve a MOP. The aim in a single-objective optimization issue is to discover one solution that maximizes the solitary objective function (unless in some special multi-modal optimization situations, when several optimum solutions are sought). Extending the concept to multi-objective optimization, it may be mistakenly believed that the aim in multi-objective optimization is to identify an optimum solution matching to the set of objectives.

Take the decision-making process involved in purchasing a car. Cars range in price from just few thousand to just a hundred thousand dollars. Let us consider two extreme hypothetical vehicles, one costing around $10,000 (solution 1) and another costing around $100,000 (solution 2), as illustrated in FigureIII.1. If the primary goal of this decision-making process is to save money, answer 1 is the best option. If this was the primary goal for all purchasers, we would only have seen one type of automobile (solution 1) on the road, and no car manufacturer would have built any costly cars. Fortunately, this decision-making process does not have a single goal. Aside from a few instances, it is assumed that an inexpensive car is likely to be less comfortable. According to the statistic, the cheapest automobile has a potential comfort level of 40%. Solution 2 is the preferred option for wealthy purchasers whose sole goal in making this purchase is comfort (with a hypothetical maximum comfort level of 90 percent, as shown in the figure). This so-called two-objective optimization issue does not have to be thought of as two separate optimization problems, the outputs of which are the two extreme solutions mentioned before.

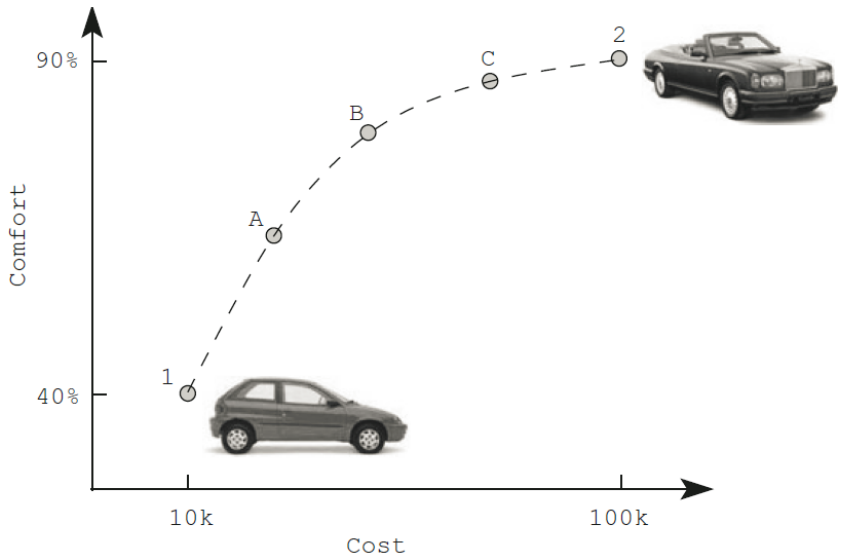


Fig.III.1. Hypothetical trade-off solutions are illustrated for a car-buying decision-making problem

Several additional options exist between these two extreme solutions, where a trade-off between expense and comfort exists. The image also depicts a variety of such solutions (solutions A, B, and C) with varying prices and satisfaction. Therefore, amongst any two such solutions, one is superior in quality of one aim, but this superiority comes exclusively at the expense of the other. All such trade-off solutions are, in this sense, optimum solutions to a MOP. Such trade-off solutions frequently give a distinct front on an objective space represented with the objective values. This is known as the Pareto-optimal front, and certain trajectories are termed Pareto-optimal trajectories.

III.1.1. Difference between Single-Objective and Multiobjective Optimization

It is obvious from the above discussion that there are several distinctions between single and multiobjective optimization problems. These latter have the following characteristics:

- Cardinality of the optimal set is usually more than one,
- There are two distinct goals of optimization, instead of one, and
- They possess two different search spaces.

We discuss each of the above properties in the following paragraphs.

To begin with, we can see from the above car-buying scenario that a multiobjective optimization with competing purposes yields a variety of Pareto optimum solutions, as opposed to the commonly held belief that a single optimal solution is associated with a single- objective optimization job. However, there are some single-objective optimization problems that have

many optimum solutions (of equal or unequal importance). Multiobjective optimization is comparable in some ways to such multi-modal optimization challenges. Nevertheless, there is a distinction to be made, which we would like to emphasize here. The input parameters of the Pareto-optimal solutions in most MOPs are identical. [7] In a multi-modal optimization problem, from the other side, from one local or global optimum solution and another. Certain choice variables have the same values in all Pareto-optimal solutions. Such a feature of the choice variables indicates that the solution is optimal.

Other decision variables take different values causing the solutions to have a trade-off in their objective values.

Secondly, unlike the sole goal of finding the optimum in a single-objective optimization, here there are two distinct goals:

- Convergence to the Pareto-optimal solutions and
- Maintenance of a set of maximally spread Pareto-optimal solutions.

In a sense, these goals are independent of each other. An optimization algorithm must have specific properties for achieving each of the goals.

One other difference between single-objective and multi-objective optimization is that in multi-objective optimization the objective functions constitute a multi-dimensional space, in addition to the usual decision variable space common to all optimization problems. This additional space is called the objective space, Z. For each solution x in the, there exists a point in the objective space, denoted by $f(x)= z = (z_1, z_2, ..., z_M)^T$. The mapping takes place between an *n* dimensional solution vector and an *M* dimensional objective vector. Figure III.2 illustrates these two spaces and a mapping between them. Although the search process of an algorithm takes place on the decision variables space, many interesting algorithms (particularly MOEAs) use the objective space information in their search operators. However, the presence of two different spaces introduces a number of interesting flexibilities in designing a search algorithm for multiobjective optimization.

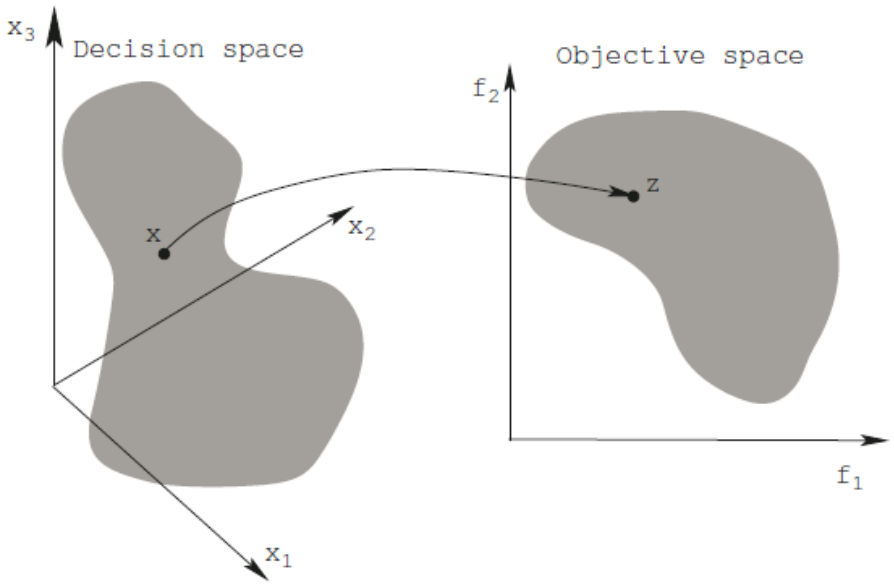


Fig.III.2. Representation of the decision variable space and the corresponding objective space.

III.2. Two Approaches to Multi-objective Optimization

Although the fundamental difference between single- and multiple-objective optimization lies in the cardinality in the optimal set, from a practical standpoint a user needs only one solution, no matter whether the associated optimization problem is single-objective or multiobjective. In the case of multiobjective optimization, the user is now in a dilemma. Which of these optimal solutions must one choose? Let us try to answer this question for the case of the car buying problem. Knowing the number of solutions that exist in the market with different trade-offs between cost and comfort, which car does one buy? This is not an easy question to answer. It involves many other considerations, such as the total finance available to buy the car, distance to be driven each day, number of passengers riding in the car, fuel consumption and cost, depreciation value, road conditions where the car is to be mostly driven, physical health of the passengers, social status and many other factors. Often, such higher-level information is nontechnical, qualitative and experience-driven. How- ever, if a set of trade-off solutions are already worked out or available, one can evaluate the pros and cons of each of these solutions based on all such non-technical and qualitative, yet still important, considerations and compare them to make a choice. Thus, in a multi-objective optimization, ideally the effort must be in finding the set of trade-off optimal solutions by considering all objectives to be important. After a set of such trade-off solutions are found, a user can then use higher-level qualitative considerations to make a choice. Therefore, we suggest the following principle for an ideal multiobjective optimization procedure:

Step 1: Find multiple trade-off optimal solutions with a wide range of values for objectives.

Step 2: Choose one of the obtained solutions using higher-level information.

Figure III.3 shows schematically the principles in an ideal multi-objective optimization procedure. In Step 1 (vertically downwards), multiple trade-off solutions are found. Thereafter, in Step 2 (horizontally, towards the right), higher-level in- formation is used to choose one of the trade-off solutions. With this procedure in mind, it is easy to realize that single-objective optimization is a degenerate case of multi-objective optimization. In the case of single-objective optimization with only one global optimal solution, Step 1 will find only one solution, thereby not requiring us to proceed to Step 2. In the case of single-objective optimization with multiple global optima, both steps are necessary to first find all or many of the global optima and then to choose one from them by using the higher-level information about the problem.

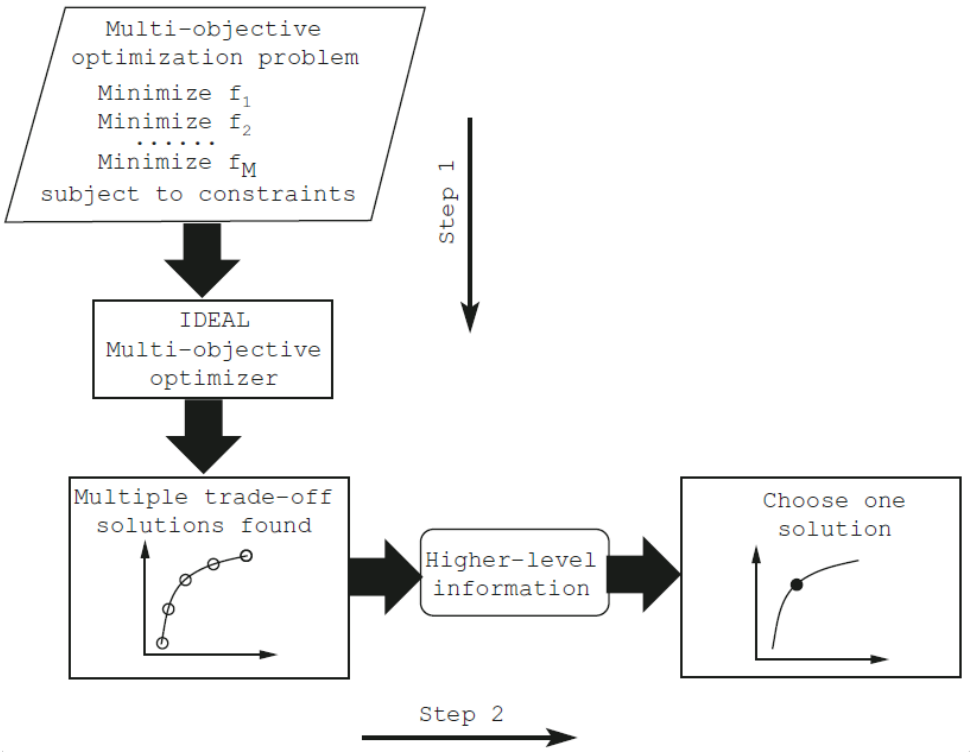


Fig.III.3. Schematic of an ideal multi-objective optimization procedure

If thought of carefully, each trade-off solution corresponds to a specific order of importance of the objectives. It is clear from Figure III.1 that solution A assigns more importance to cost than to comfort. On the other hand, solution C assigns more im- portance to comfort than to cost. Thus, if such a relative preference factor among the objectives is known for a specific problem, there is no need to follow the above principle for solving a MOP. A simple method would be to form a composite objective function as the weighted sum of the objectives, where

a weight for an objective is proportional to the preference factor assigned to that particular objective. This method of scalarizing an objective vector into a single composite objective function converts the MOP into a single-objective optimization problem. When such a composite objective function is optimized, in most cases it is possible to obtain one particular trade-off solution. This procedure of handling MOPs is much simpler, though still being more subjective than the above ideal procedure. We call this procedure a preference-based multi-objective optimization. A schematic of this procedure is shown in Figure III.4.

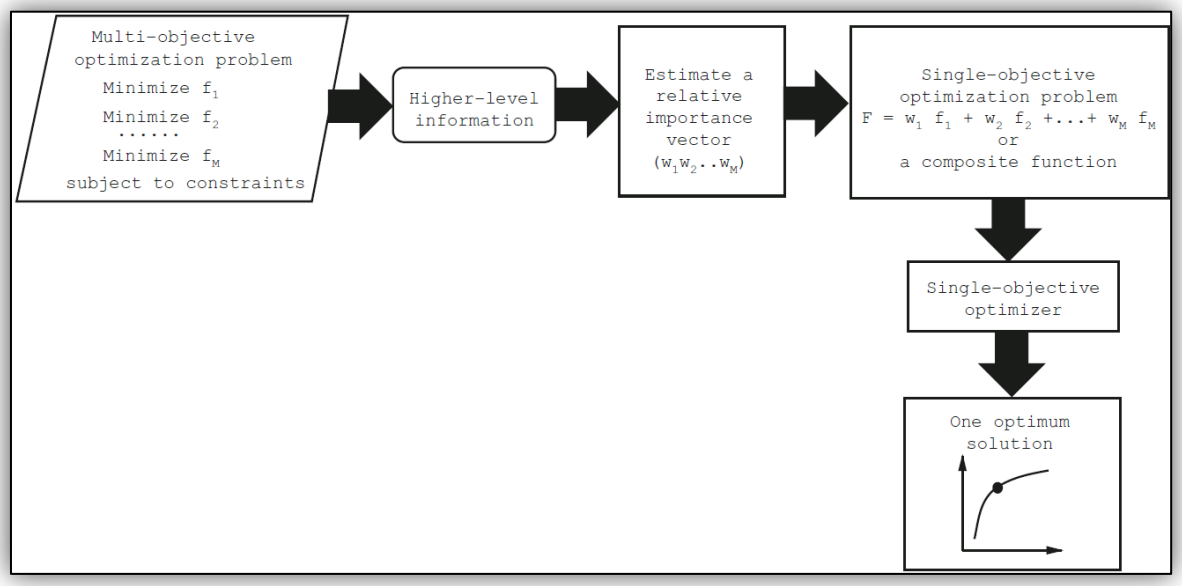


Fig.III.4 Schematic of a preference-based multi-objective optimization procedure

Based on the higher-level information, a preference vector w is first chosen. Thereafter, the preference vector is used to construct the composite function, which is then optimized to find a single trade-off optimal solution by a single-objective optimization algorithm. Although not often practiced, the procedure can be used to find multiple trade-off solutions by using a different preference vector and repeating the above procedure.

It is important to appreciate that the trade-off solution obtained by using the preference-based strategy is largely sensitive to the relative preference vector used in forming the composite function. A change in this preference vector will result in a (hopefully) different trade-off solution. Besides this difficulty, it is intuitive to realize that finding a relative preference vector itself is highly subjective and not straightforward. This requires an analysis of the non-technical, qualitative and experience-driven information to find a quantitative relative preference vector. Without any knowledge of the likely trade-off solutions, this is an

even more difficult task. Classical multi-objective optimization methods which convert multiple objectives into a single objective by using a relative preference vector of objectives work according to this preference-based strategy. Unless a reliable and accurate preference vector is available, the optimal solution obtained by such methods is highly subjective to the particular user.

The ideal multi-objective optimization procedure suggested earlier is less subjective. In Step 1, a user does not need any relative preference vector information. The task there is to find as many different trade-off solutions as possible. Once a well-distributed set of trade-off solutions is found, Step 2 then requires certain problem information in order to choose one solution. It is important to mention that in Step 2, the problem information is used to evaluate and compare each of the obtained trade-off solutions. In the ideal approach, the problem information is not used to search for a new solution; instead, it is used to choose one solution from a set of already obtained trade-off solutions. Thus, there is a fundamental difference in using the problem information in both approaches. In the preference-based approach, a relative preference vector needs to be supplied without any knowledge of the possible consequences. However, in the proposed ideal approach, the problem information is used to choose one solution from the obtained set of trade-off solutions. We argue that the ideal approach in this matter is more methodical, more practical, and less subjective. At the same time, we highlight the fact that if a reliable relative preference vector is available to a problem, there is no reason to find other trade-off solutions. In such a case, a preference-based approach would be adequate.

In the next section, we make the above qualitative idea of multi-objective optimization more quantitative.

III.3. Non-dominated Solutions and Pareto-Optimal Solutions

Most multi-objective optimization algorithms use the concept of dominance in their search. Here, we define the concept of dominance and related terms and present a number of techniques for identifying dominated solutions in a finite population of solutions.

III.3.1. Special Solutions

We first define some special solutions which are often used in multi-objective optimization algorithms.

III.3.1.1. Ideal Objective Vector

For each of the M conflicting objectives, there exists one different optimal solution. An objective vector constructed with these individual optimal objective values constitutes the ideal objective vector.

Definition III.1: The mth component of the ideal objective vector $\mathbf{z}^*$ is the constrained minimum solution of the following problem:

$$\left.\begin{array}{l}\text{Minimize } f_m(\mathrm{x}) \\ \text{subject to } \mathrm{x} \in \mathrm{S}\end{array}\right\} \qquad \text{(III.1)}$$

Thus, if the minimum solution for the mth objective function is the decision vector $\mathbf{x}^{*(m)}$ with function value f_m^*, the ideal vector is as follows:

$$\mathbf{z}^* = \mathbf{f}^* = (f_1^*, f_2^*, \ldots, f_M^*)^{\mathrm{T}}.$$

In general, the ideal objective vector ($\mathbf{z}^*$) corresponds to a non-existent solution (Figure III.5). This is because the minimum solution of equation (III.1) for each objective function need not be the same solution. The only way an ideal objective vector corresponds to a feasible solution is when the minimal solutions to all objective functions are identical. In this case, the objectives are not conflicting to each other and the minimum solution to any objective function would be the only optimal solution to the MOP. Although the ideal objective vector is usually non-existent, it is also clear from Figure III.5 that solutions closer to the ideal objective vector are better. Moreover, many algorithms require the knowledge of the lower bound on each objective function to normalize objective values in a common range.

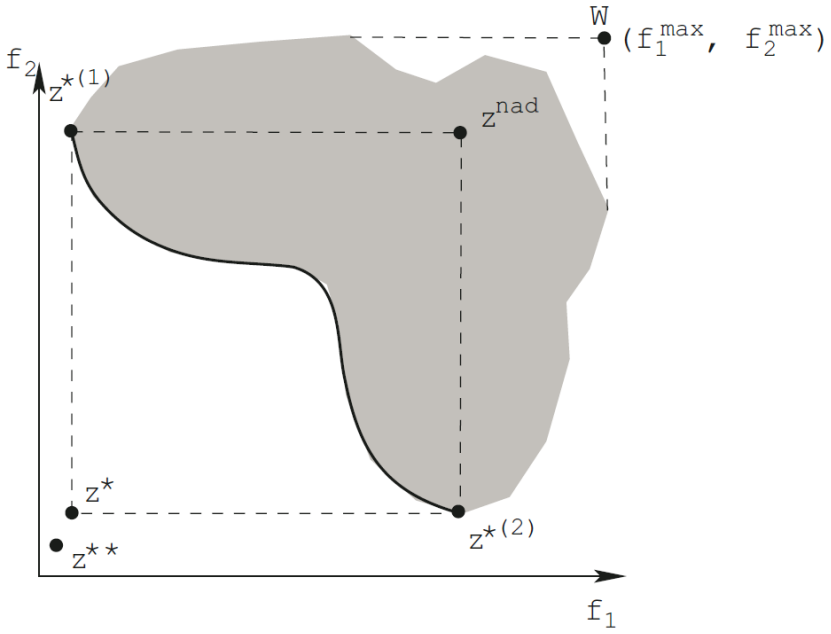


Fig.III.5. The ideal, utopian and nadir objective vectors

III.3.1.2. Utopian Objective Vector

The ideal objective vector denotes an array of the lower bound of all objective functions. This means that for every objective function there exists at least one solution in the feasible search space sharing an identical value with the corresponding element in the ideal solution. Some algorithms may require a solution which has an objective value strictly better than (and not equal to) that of any solution in the search space. For this purpose, the utopian objective vector is defined as follows:

Definition III.2: A utopian objective vector $\mathbf{z}^{**}$ has each of its components marginally smaller than that of the ideal objective vector, or $\mathbf{z}_i^{**} = \mathbf{z}_i^{*} - \varepsilon_i$ with $\varepsilon_i > 0$ for all i = 1, 2, ..., *M*.

Figure III.5 shows a utopian objective vector. Like the ideal objective vector, the utopian objective vector also represents a non-existent solution.

III.3.1.3. Nadir Objective Vector

Unlike the ideal objective vector which represents the lower bound of each objective in the entire feasible search space, the nadir objective vector $\mathbf{z}^{nad}$ represents the upper bound of each objective in the entire Pareto-optimal set, and not in the entire search space. A nadir objective vector must not be confused with a vector of objectives (marked as “W” in Fig.III.5) found by using the worst feasible function values f max in the entire search space. The nadir objective vector may represent an existent or a non-existent solution, depending on the convexity and

continuity of the Pareto-optimal set. In order to normalize each objective in the entire range of the Pareto-optimal region, the knowledge of nadir and ideal objective vectors can be used as follows:

$$f_i^{\text{norm}} = \frac{f_i - z_i^*}{z_i^{\text{nad}} - z_i^*}. \qquad \text{(III.2)}$$

III.3.2. Concept of Domination

Most multi-objective optimization algorithms use the concept of domination. In these algorithms, two solutions are compared on the basis of whether one dominates the other or not. We will describe the concept of domination in the following paragraph.

We assume that there are M objective functions. In order to cover both minimization and maximization of objective functions, we use the operator $\lhd$ between two solutions i and j as $i \lhd j$ to denote that solution i is better than solution j on a particular objective. Similarly, $i \rhd j$ for a particular objective implies that solution i is worse than solution j on this objective. For example, if an objective function is to be minimized, the operator $\lhd$ would mean the "$<$" operator, whereas if the objective function is to be maximized, the operator $\lhd$ would mean the "$>$" operator. The following definition covers mixed problems with minimization of some objective functions and maximization of the rest of them.

Definition III.3. A solution $\mathbf{x}^{(1)}$ is said to dominate the other solution $\mathbf{x}^{(2)}$ if both conditions 1 and 2 are true:

1. The solution $\mathbf{x}^{(1)}$ is no worse than $\mathbf{x}^{(2)}$ in all objectives, or $f_j(\mathbf{x}^{(1)}) \not\lhd f_j(\mathbf{x}^{(2)})$ for all $j = 1, 2, \ldots, M$.
2. The solution $\mathbf{x}^{(1)}$ is strictly better than $\mathbf{x}^{(2)}$ in at least one objective, or $f_{\bar{j}}(\mathbf{x}^{(1)}) \lhd f_{\bar{j}}(\mathbf{x}^{(2)})$ for at least one $\bar{j} \in \{1, 2, \ldots, M\}$.

If either of these conditions is violated, the solution $\mathbf{x}^{(1)}$ does not dominate the solution $\mathbf{x}^{(2)}$. If $\mathbf{x}^{(1)}$ dominates the solution $\mathbf{x}^{(2)}$ (or mathematically $\mathbf{x}^{(1)} \prec \mathbf{x}^{(2)}$), it is also customary to write any of the following:

- $\mathbf{x}^{(2)}$ is dominated by $\mathbf{x}^{(1)}$
- $\mathbf{x}^{(1)}$ is non-dominated by $\mathbf{x}^{(2)}$ or
- $\mathbf{x}^{(1)}$ is non-inferior to $\mathbf{x}^{(2)}$.

Let us consider a two-objective optimization problem with five different solutions shown in the objective space, as illustrated in Figure III.6a. Let us also assume

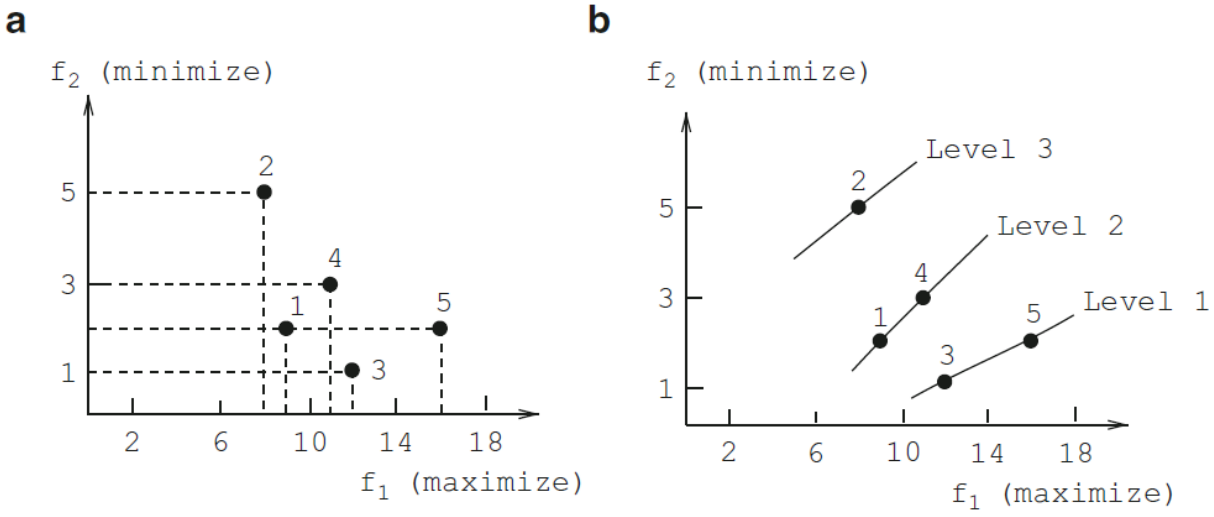


Fig.III.6. A set of five solutions and the corresponding non-dominated fronts

that the objective function 1 needs to be maximized while the objective function 2 needs to be minimized. Five solutions with different objective function values are shown in this figure. Since both objective functions are of importance to us, it is usually difficult to find one solution which is best with respect to both objectives. However, we can use the above definition of domination to decide which solution is better among any two given solutions in terms of both objectives. For example, if solutions 1 and 2 are to be compared, we observe that solution 1 is better than solution 2 in objective function 1 and solution 1 is also better than solution 2 in objective function 2. Thus, both of the above conditions for domination are satisfied and we may write that solution 1 dominates solution 2. We take another instance of comparing solutions 1 and 5. Here, solution 5 is better than solution 1 in the first objective and solution 5 is no worse (in fact, they are equal) than solution 1 in the second objective. Thus, both the above conditions for domination are also satisfied and we may write that solution 5 dominates solution 1.

It is intuitive that if a solution $\mathbf{x}^{(1)}$ dominates another solution $\mathbf{x}^{(2)}$, the solution $\mathbf{x}^{(1)}$ is better than $\mathbf{x}^{(2)}$ in the parlance of multi-objective optimization. Since the concept of domination allows a way to compare solutions with multiple objectives, most multi-objective optimization

methods use this domination concept to search for non-dominated solutions.

III.3.3. Properties of Dominance Relation

Definition 15.3 defines the dominance relation between any two solutions. There are three possibilities that can be the outcome of the dominance check between two solutions 1 and 2. That is (i) solution 1 dominates solution 2, (ii) solution 1 gets dominated by solution 2, or (iii) solutions 1 and 2 do not dominate each other. Let us now discuss the different binary relation properties [8] of the dominance operator.

- *Reflexive*. The dominance relation is *not reflexive*, since any solution p does not dominate itself according to Definition III.3. The second condition of dominance relation in Definition III.3 does not allow this property to be satisfied.
- *Symmetric*. The dominance relation is also no*t symmetric*, because $p \leq q$ does not imply $q \leq p$. In fact, the opposite is true. That is, if p dominates q, then q does not dominate p. Thus, the dominance relation is asymmetric.
- *Antisymmetric*. Since the dominance relation is not symmetric, it cannot be anti-symmetric as well.
- *Transitive*. The dominance relation is transitive. This is because if $p \leq q$ and $q \leq r$, then $p \leq r$.

There is another interesting property that the dominance relation possesses. If solution p does not dominate solution q, this does not imply that q dominates p.

In order for a binary relation to qualify as an ordering relation, it must be at least transitive [9]. Thus, the dominance relation qualifies as an ordering relation. Since the dominance relation is not reflexive, it is a strict partial order. In general, if a relation is reflexive, antisymmetric and transitive, it is loosely called a partial order and a set on which a partial order is defined is called a partially ordered set. However, it is important to note that the dominance relation is not reflexive and is not antisymmetric. Thus, the dominance relation is not a partial-order relation in its general sense. The dominance relation is only a strict partial-order relation.

III.3.4. Pareto Optimality

Continuing with the comparisons in the previous section, let us compare solutions 3 and 5 in Fig.III.6, because this comparison reveals an interesting aspect. We observe that solution 5 is better than solution 3 in the first objective, while solution 5 is worse than solution 3 in the second objective. Thus, the first condition is not satisfied for both of these solutions. This simply suggests that we cannot conclude that solution 5 dominates solution 3, nor can we say that solution 3 dominates solution 5. When this happens, it is customary to say that solutions 3 and 5 are non-dominated with respect to each other. When both objectives are important, it cannot be said which of the two solutions 3 and 5 is better.

For a given finite set of solutions, we can perform all possible pair-wise comparisons and find which solution dominates which and which solutions are non-dominated with respect to each other. At the end, we expect to have a set of solutions, any two of which do not dominate each other. This set also has another property. For any solution outside of this set, we can always find a solution in this set which will dominate the former. Thus, this particular set has a property of dominating all other solutions which do not belong to this set. In simple terms, this means that the solutions of this set are better compared to the rest of the solutions. This set is given a special name. It is called the non-dominated set for the given set of solutions. In the example problem, solutions 3 and 5 constitute the non-dominated set of the given set of five solutions. Thus, we define a set of non-dominated solutions as follows.

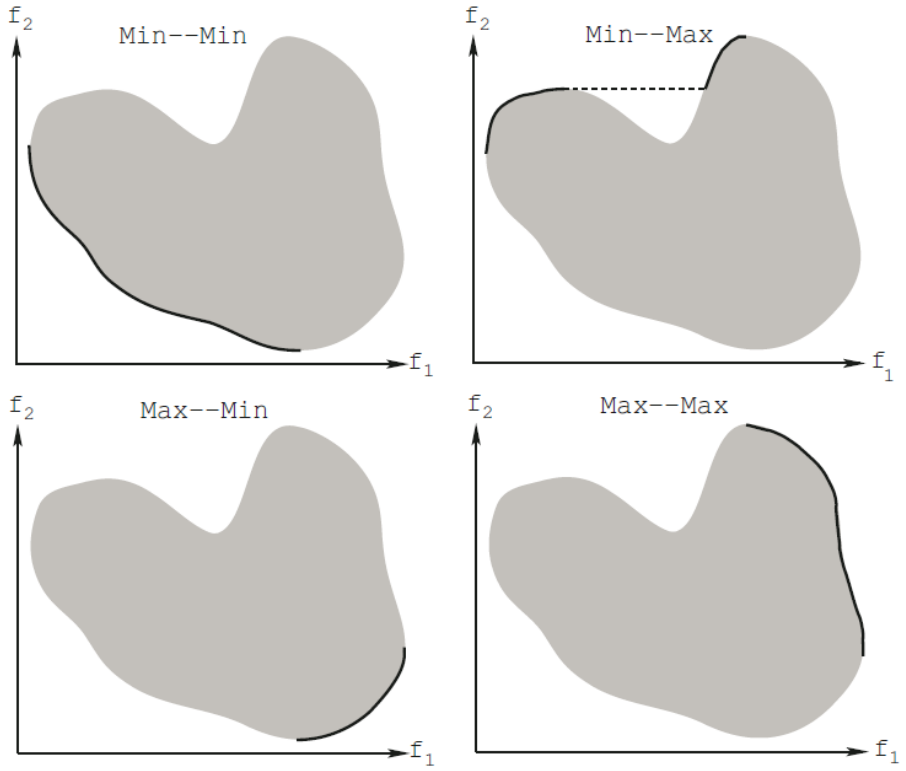


Fig.III.7. Pareto-optimal solutions are marked with continuous curves for four combinations of two types of objectives.

Definition III.4 (Non-dominated set). Among a set of solutions P, the non-dominated set of solutions P' are those that are not dominated by any member of the set P.

When the set P is the entire search space, or $P = \mathcal{S}$, the resulting non-dominated set P' is called the *Pareto-optimal set*. Figure III.7 marks the Pareto-optimal set with continuous curves for four different scenarios with two objectives. Each objective can be minimized or maximized. In the top-left panel, the task is to minimize both objectives f_1 and f_2. The solid curve marks the Pareto-optimal solution set. If f_1 is to be minimized and f_2 is to be maximized for a problem having the same search space, the resulting Pareto-optimal set is different and is shown in the top-right panel. Here, the Pareto-optimal set is a union of two disconnected Pareto-optimal regions.

Similarly, the Pareto-optimal sets for two other cases—(maximizing f_1, minimizing f_2) and (maximizing f_1, maximizing f_2)—are shown in the bottom-left and bottom-right panels, respectively. In any case, the Pareto-optimal set always consists of solutions from a particular edge of the feasible search region.

It is important to note that an MOEA can be easily used to handle all of the above cases by simply using the domination definition. However, to avoid any confusion, most applications use the duality principle to convert a maximization problem into a minimization problem and treat every problem as a combination of minimizing all objectives. Like global and local optimal solutions in the case of single-objective optimization, there could be global and local Pareto-optimal sets in multi-objective optimization.

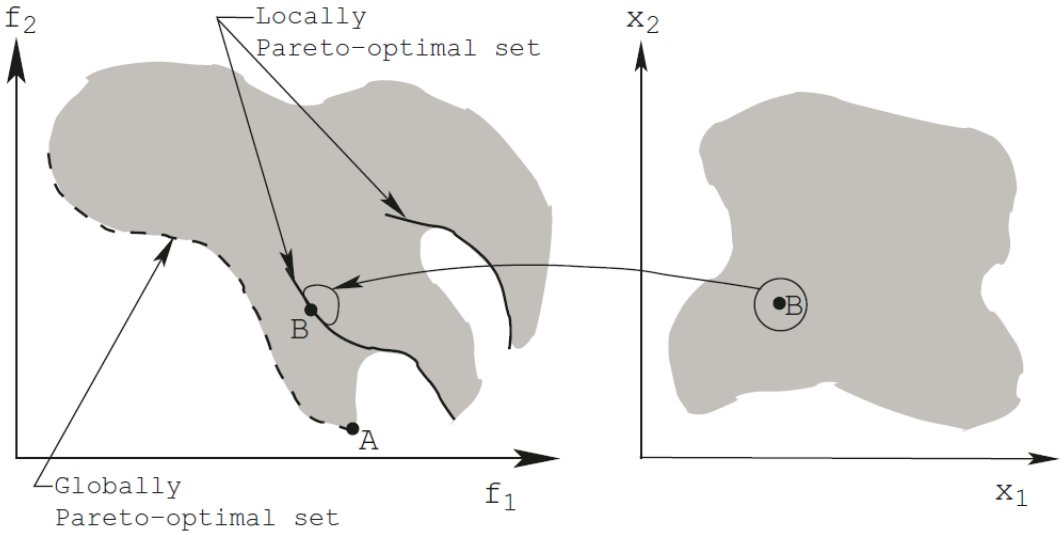


Fig.III.8 Locally and globally Pareto-optimal solutions

Definition III.5: (Globally Pareto-optimal set). The non-dominated set of the entire feasible search space S is the globally Pareto-optimal set.

Definition III.6: If for every member *x* in a set *P* there exists no solution y (in the neighborhood of x such that $\|y - x\|_\infty$, where is a small positive number) dominating any member of the set *P*, then solutions belonging to the set *P* constitute a locally Pareto-optimal set.

Figure III.8 shows two locally Pareto-optimal sets (marked by continuous curves). When any solution (say "B") in this set is perturbed locally in the decision variable space, no solution can be found dominating any member of the set. It is interesting to note that for continuous search space problems, the locally Pareto- optimal solutions need not be continuous in the decision variable space and the above definition will still hold good. Zitzler (1999) added a neighborhood constraint on the objective space in the above definition to make it more generic. By the above definition, it is also true that a globally Pareto-optimal set is also a locally Pareto optimal set.

III.3.5. Procedure for Finding Non-dominated Solutions

Finding the non-dominated set of solutions from a given set of solutions is similar in principle to finding the minimum of a set of real numbers. In the latter case, when two numbers are compared to identify the smaller number, a '$<$' relation operation is used. In the case of finding the non-dominated set, the dominance relation can be used to identify the better of two given solutions. Here, we discuss one simple procedure for finding the non-dominated set (we call here the best non-dominated front). Many MOEAs require to find the best non-dominated solutions of a population and some MOEAs require to sort a population according to different non-domination levels. We present one algorithm for each of the tasks.

III.3.5.1. Finding the Best Non-dominated Front

In this approach, every solution from the population is checked with a partially filled population for domination. To start with, the first solution from the population is kept in an empty set P'. Thereafter, each solution i (the second solution onwards) is compared with all members of the set P', one by one. If the solution i dominates any member of P', then that solution is removed from P'. In this way non-members of the non-dominated solutions get deleted from P'. Otherwise, if solution i is dominated by any member of P', the solution i is ignored. If solution i is not dominated by any member of P', it is entered in P'. This is how the set P' grows with non-dominated solutions. When all solutions of the population are checked, the remaining members of P' constitute the non-dominated set.

Identifying the non-dominated set:

Step 1: Initialize $P' = 1$. Set solution counter $i = 2$.

Step 2: Set $j = 1$.

Step 3: Compare solution i with j from P' for domination.

Step 4: If i dominates j, then delete the jth member from P' or else update $P' = P' \setminus \{P'(j)\}$.

If $j < P'$, increment j by one and then go to **Step 3**. Otherwise, go to **Step 5.** Alternatively, if the jth member of P' dominates i, increment i by one and then go to **Step 2.**

Step 5 Insert i in P' or update $P' = P' \cup \{i\}$. If $i < N$, increment i by one and go to **Step 2**. Otherwise, stop and declare P' as the non-dominated set.

Here, we observe that the second element of the population is compared with only one solution P', the third solution with at most two solutions of P', and so on. This requires a maximum of

$1 + 2 + \ldots + (N - 1)$ or $N(N\ 1)/2$ domination checks. This computation is also $O(MN^2)$. It is interesting to note that the size of P' may not always increase (dominated solutions will get deleted from P') and not every solution in the population may be required to be checked with all solutions in the current P' set (the solution may get dominated by a solution of P'). Thus, the actual computational complexity may be smaller than the above estimate.

Another study [10] suggested a binary-search-like algorithm for finding the best non-dominated front with a complexity $O(N(\log N)^{M-2})$ for $M \geq 4$ and $O(N \log N)$ for $M = 2$ and 3.

III.3.5.2. A Non-dominated Sorting Procedure

Using the above procedure, each front can be identified with at most $O(MN^2)$ computations. In certain scenarios, this procedure may demand more than $O(MN^2)$ computational effort for the overall non-dominated sorting of a population. Here, we suggest a completely different procedure which uses a better bookkeeping strategy requiring $O(MN^2)$ overall computational complexity.

First, for each solution we calculate two entities: (i) *domination count* n_i, the number of solutions which dominate the solution i, and (ii) S_i, a set of solutions which the solution i dominates. This requiers $O(MN^2)$ comparisons. At the end of this procedure, all solutions in the first non-dominated front will have their domination count as zero. Now, for each of these solutions (each solution i with $n_i = 0$), we visit each member (j) of its set S_i and reduce its domination count by one. In doing so, if for any member j the domination count becomes zero, we put it in a separate list P'. After such modifications on S_i are performed for each i with $n_i = 0$, all solutions of P' would belong to the second non-dominated front. The above procedure can be continued with each member of P' and the third non-dominated front can be identified. This process continues until all solutions are classified.

An O(MN^2) non-dominated sorting algorithm:

Step 1: For each $i \in P$, $n_i = 0$ and initialize $S_i = \emptyset$. For all j $\neq i$ and $j \in P$, perform **Step 2** and then proceed to **Step 3**.

Step 2: If $i \leq j$, update $S_p = S_p \cup \{j\}$. Otherwise, if $j \leq i$, set $n_i = n_i + 1$.

Step 3: If $n_i = 0$, keep i in the first non-dominated front P_1 (we called this set P' in the above paragraph). Set a front counter $k = 1$.

Step 4: While $P_k = \emptyset$, perform the following steps.

Step 5: Initialize $Q = \emptyset$ for storing next non-dominated solutions. For each $I \in P_k$ and for each $j \in S_i$,

Step 5a: Update $n_j = n_j$ - 1.

Step 5b: If $n_j = 0$, keep j in Q, or perform $Q = Q \cup \{j\}$.

Step 6: Set $k = k + 1$ and $P_k = Q$. Go to **Step 4.**

Steps 1–3 find the solutions in the first non-dominated front and require O(MN^2) computational complexity. Steps 4–6 repeatedly find higher fronts and require at most O(N^2) comparisons, as argued below. For each solution i in the second- or higher level of non-domination, the domination count n_i can be at most N 1. Thus, each solution i will be visited at most N - 1 times before its domination count becomes zero. At this point, the solution is assigned a particular non-domination level and will never be visited again. Since there are at most N - 1 such solutions, the complexity of identifying second and more fronts is O(N^2). Thus, the overall complexity of the procedure is O(MN^2). It is important to note that although the time complexity has reduced to O(MN^2), the storage requirement has increased to O(N^2).

When the above procedure is applied to the five solutions of Figure III.6a, we obtain three non-dominated fronts as shown in Figure III.6b. From the dominance relations, the solutions 3 and 5 are the best, followed by solutions 1 and 4. Finally, solution 2 belongs to the worst non-dominated front. Thus, the ordering of solutions in terms of their non-domination level is as follows: ((3,5), (1,4), (2)). A study (Jensen 2003b) suggested a divided-and-conquer method to reduce the complexity of sorting to O($N \log^{M-1} N$).

III.4. Some Approaches to Multi-objective Optimization

In this section, we briefly mention two commonly used classical multi-objective optimization methods and thereafter present a commonly used EMO method.

III.4.1 Classical Method: Weighted-Sum Approach

The weighted-sum method, as the name suggests, scalarizes a set of objectives into a single objective by pre-multiplying each objective with a user-supplied weight. This method is the simplest approach and is probably the most widely used classical approach. If we are faced with the two objectives of minimizing the cost of a product and minimizing the amount of wasted material in the process of fabricating the product, one naturally thinks of minimizing a weighted sum of these two objectives. Although the idea is simple, it introduces a not-so-simple question. What values of the weights must one use? Of course, there is no unique answer to this question. The answer depends on the importance of each objective in the context of the problem and a scaling factor. The scaling effect can be avoided somewhat by normalizing the objective functions. After the objectives are normalized, a composite objective function $F(\mathrm{x})$ can be formed by summing the weighted normalized objectives and the problem is then converted to a single-objective optimization problem as follows:

$$\left.\begin{array}{lll} \text{Minimize } F(\mathbf{x}) = \sum_{m=1}^{M} w_m f_m(\mathbf{x}) & & \\ \text{subject to } g_j(\mathbf{x}) \geq 0, & j = 1,2,\ldots,J \\ h_k(\mathbf{x}) = 0, & k = 1,2,\ldots,K \\ x_i^{(L)} \leq x_i \leq x_i^{(U)}, & i = 1,2,\ldots,n. \end{array}\right\} \qquad \text{(III.3)}$$

Here, w_m ($\in [0, 1]$) is the weight of the mth objective function. Since the minimum of the above problem does not change if all weights are multiplied by a constant, it is the usual practice to choose weights such that their sum is one, or $\sum_{m=1}^{M} w_m = 1$.

Mathematically oriented readers may find a number of interesting theorems regarding the relationship between the optimal solution of the above problem to the true Pareto-optimal solutions in classical texts [9,11].

Let us now illustrate how the weighted-sum approach can find Pareto-optimal solutions of the original problem. For simplicity, we consider the two-objective problem shown in Figure III.9.

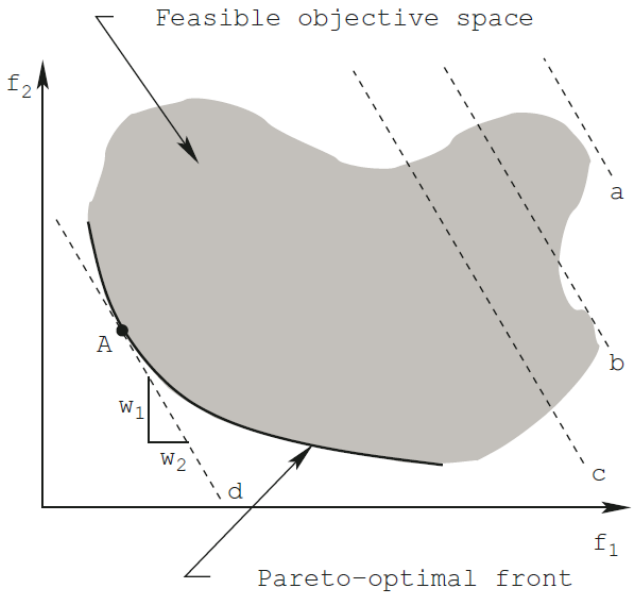


Fig.III.9 Illustration of the weighted-sum approach on a convex Pareto-optimal front

The feasible objective space and the corresponding Pareto-optimal solution set are shown. With two objectives, there are two weights w_1 and w_2, but only one is independent. Knowing any one, the other can be calculated by simple subtraction. It is clear from the figure that a choice of a weight vector corresponds to a pre-destined optimal solution on the Pareto-optimal front, as marked by the point A. By changing the weight vector, a different Pareto-optimal point can be obtained. However, there are a couple of difficulties with this approach:

1. A uniform choice of weight vectors does not necessarily find a uniform set of Pareto-optimal solutions on the Pareto-optimal front [1].
2. The procedure cannot be used to find Pareto-optimal solutions which lie on the non-convex portion of the Pareto-optimal front.

The former issue makes it difficult for the weighted-sum approach to be applied reliably to any problem in order to find a good representative set of Pareto-optimal solutions. The latter issue arises due to the fact that a solution lying on the non- convex Pareto-optimal front can never be the optimal solution of the problem given in Equation (III.3).

III.4.2. Classical Method: ε-Constraint Method

In order to alleviate the difficulties faced by the weighted-sum approach in solving problems having non-convex objective spaces, the **ε** -constraint method is used. In 1971, Haimes suggested reformulating the MOP by just keeping one of the objectives and restricting the rest of the objectives within user-specified values [12]. The modified problem is as follows:

$$\left.\begin{array}{ll} \text{Minimize } f_\mu(\mathbf{x}) & \\ \text{subject to } f_m(\mathbf{x}) \le \varepsilon_m, & m = 1,2,\ldots,M \text{ and } m \ne \mu \\ g_j(\mathbf{x}) \ge 0, & j = 1,2,\ldots, \\ h_k(\mathbf{x}) = 0, & k = 1,2,\ldots,K \\ x_i^{(L)} \le x_i \le x_i^{(U)}, & i = 1,2,\ldots,n. \end{array}\right\} \qquad \text{(III.4)}$$

In the above formulation, the parameter ε_m represents an upper bound of the value of f_m and need not necessarily mean a small value close to zero.

Let us say that we retain f_2 as an objective and treat f_1 as a constraint: $f_1(\mathrm{x}) \le \varepsilon_1$.

Figure III.10 shows four scenarios with different ε_1 values. Let us consider the third scenario with $\varepsilon_1 = \varepsilon_1^c$ first. The resulting problem with this constraint divides the original feasible objective space into two portions, $f_1 \le \varepsilon_1^c$ and $f_1 > \varepsilon_1^c$. The left portion becomes the feasible solution of the resulting problem stated in Equation (III.4). Now, the task of the resulting problem is to find the solution which minimizes this feasible region. From Figure (III.10), it is clear that the minimum solution is C. In this way, intermediate Pareto-optimal solutions can be obtained in the case of non-convex objective space problems by using the **ε** -constraint method.

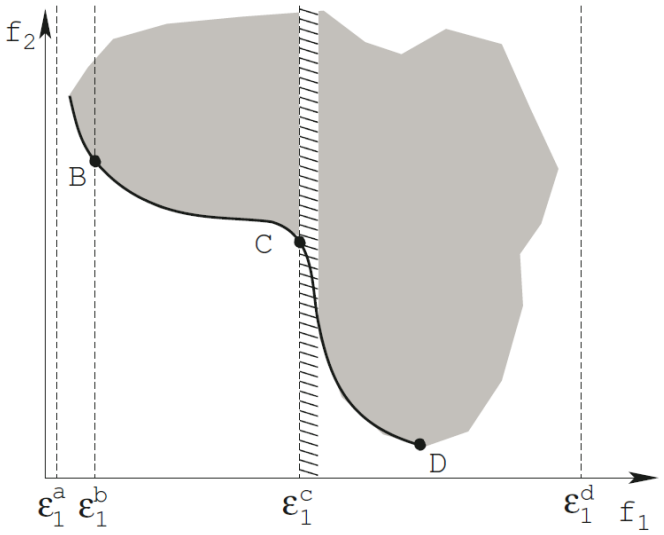


Fig.III.10. The **ε** -constraint method

One of the difficulties of this method is that the solution to the problem stated in Equation (III.4) largely depends on the chosen $\boldsymbol{\varepsilon}$ vector. Let us refer to Figure III.10 again.

Instead of choosing $\boldsymbol{\varepsilon}_1^c$, if $\boldsymbol{\varepsilon}_1^a$ is chosen, there exists no feasible solution to the stated problem. Thus, no solution would be found. On the other hand, if $\boldsymbol{\varepsilon}_1^d$ is used, the entire search space is feasible. The resulting problem has the minimum at D. Moreover, as the number of objectives increases, there exist more elements in the $\boldsymbol{\varepsilon}$ vector, thereby requiring more information from the user.

III.4.3 Evolutionary Multi-objective Optimization (EMO) Method

Over the years, a number of multiobjective EAs emphasizing non-dominated solutions in an EA population have been suggested. In this section, we shall describe one state-of-the-art algorithm popularly used in EMO studies.

Genetic algorithms (GAs) are general-purpose search algorithms widely employed in different fields of science and engineering as both optimization algorithms and scientific models of evolution. Theoretical foundations and the success in first practical applications have stimulated the study on GAs and new classes of algorithms have been proposed in the literature. In particular, multiobjective GAs (MGA) are gaining the attention of the scientific community as powerful search algorithms for complex problems. Applications of GAs in geomorphology are quite recent. The first applications can be dated back to the late 1990s, whereas applications of multiobjective versions are still more recent.

Elitist Non-dominated Sorting Genetic Algorithm (NSGA-II) nowadays represents the most widely used MGA in engineering and scientific fields and is still widely considered the state-of-the-art for practical applications. For these reasons, it is here in chosen as the reference MGA and illustrated in the following section.

III.4.3.1. Elitist Non-dominated Sorting GA (NSGA-II)

III.4.3.1.1. Introduction to Genetic algorithms

Genetic algorithms (GAs) have proved to be an effective and robust support tool for the prediction and modeling of complex phenomena. GAs belongs to the broader family of evolutionary algorithms (EAs) and can be considered as both artificial models of natural evolution and general-purpose search algorithms.

The initial studies on GAs go back to the 1960s, when a growing number of researchers began to consider natural systems as a source of inspiration for the development of optimization algorithms for engineering problems. Among these, John Holland, who is universally recognized as the father of GAs, was interested in the principles governing the evolution of adaptive natural systems, speculating that competition and innovation were the key mechanisms through which individuals acquire the ability to adapt themselves to the environment [13].

III.4.3.1.2. The Holland's Model

The GA proposed by Holland in 1975 is an iterative algorithm that operates on a population of N bit strings of prefixed length l (l, $N \in$ N) where each string (genotype) is the binary encoding of a candidate solution (phenotype) of a particular research problem [14]. For example, the genotype can encode specific values of a set of parameters $\pi = \{p_i \, \mathrm{l} j = 1, 2, \ldots, n\}$ of a given simulation model, where each parameter p_j is allowed to vary into a predefined range $[\alpha_j, \beta_j] \subset$ R. Note that the cardin-ality of the set of binary strings of length l grows exponentially with l, having 2^l elements. This set represents the GA search space, that is, the space that the GA needs to explore to solve the research problem.

The objective function, f, assigns a fitness value, $f_i = f(g_i)$, to each genotype $g_i (i = 1, \ldots, N)$ of the GA. To determine such value, the fitness function decodes the genotype in the corresponding phenotype and tests it on the problem producing a value, generally a real number, representing its ability to solve the problem. The graph of fitness values plotted against the search space points is called a fitness landscape.

The original Holland's model is today known as a generational scheme model, because each iteration (also called generation) replaces all the N individuals in the population with as many offspring, whereas the selection method is known as proportional selection because it selects individuals to be reproduced with a probability which is proportional to their fitness. Holland used genetic operators such as (single-point) crossover, mutation, and inversion. The inversion operator has been, however, rarely used in practical applications and rarely considered in theoretical studies. Therefore, in-version is not discussed in the following sections.

III.4.3.1.3. Proportional selection

Proportionally to their fitness values, f_i, the probabilities $p_{selection,i}$ defined as:

$$p_{\text{selection},i} = \frac{f_i}{\sum_{j=1}^{N} f_j}$$

are associated to the genotypes g_i and used to construct a sort of roulette of probability which is used in the selection process. Let us consider an example: if the population is com-posed by the $n = 4$ individuals A_1, A_2, A_3 and A_4, with probability of selection $p_{selection,1} = 0.12$, $p_{selection,2} = 0.18$, $p_{selection,3} = 0.3$, and $p_{selection,4} = 0.4$, respectively, the corresponding roulette will have the form shown in Figure III.11. The selection operator generates a random number $c \in [0, 1]$ and selects the individual associated with the roulettes' portion containing the value c. For instance, if $c = 0.78$, the individual A_4 is selected, because c falls within the range [0.6, 1]. When an individual is selected, a copy is made and inserted into the so-called mating pool. Once the mating pool is filled with exactly N copies of individuals of the population $P(t)$, members of the new population P (t + 1) are obtained as their offspring through the application of genetic operators. The selection operator, therefore, determines which individuals of the old population have the chance to generate offspring. As individuals with higher fitness are favored in the selection process, having on average a higher number of copies in the mating pool, the selection operator plays the role of Darwinian Natural Selection within the GAs context.

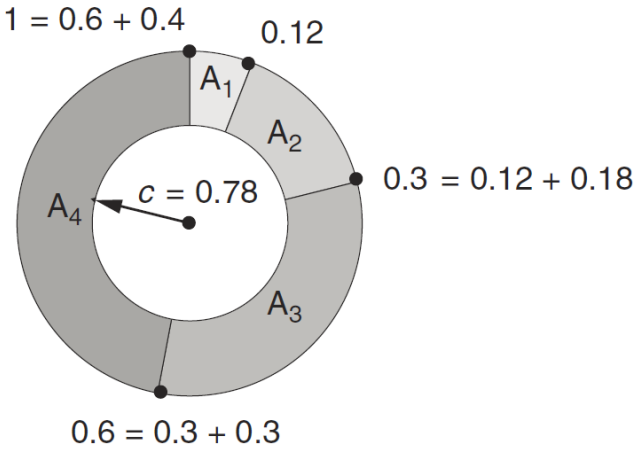


Fig.III.11. Example of proportional selection. The four individuals A_1, A_2, A_3, and A_4 hold portions of the roulette proportionally to their selection probabilities, which are set to 0.12, 0.18, 0.3, and 0.4, respectively. In the example, the selection operator generates the random number $c = 0.78$ and the individual A_4 is selected.

III.4.3.1.4. Crossover and mutation

Regarding crossover, two parent individuals are randomly chosen from the mating pool and a cutting, or crossover, point selected. Portions of the genotype are then exchanged, generating two offspring. Figure III.12 (a) shows an example of cross-over between two binary genotypes. The crossover operator is applied according to a prefixed probability, $p_{crossover}$, for a total of $N/2$ times, in order to obtain N offspring. When the crossover is not applied, offspring coincide with parents. Note that, as the selection operator plays in the GA framework the role of natural selection, crossover is a metaphor of sexual reproduction in which genetic material of offspring results in a recombination of those of the parents.

Once N offspring are obtained by crossover, mutation is applied. According to a prefixed and usually small probability, $p_{mutation}$, the bit value of each individual is simply changed from 0 to 1, or vice versa (i.e., from 1 to 0 – see Figure III12 (b)). The mutation operator represents the genetic phenomenon of the rare variation of genotype's elements in living beings during evolution.

After crossover and mutation are applied to the individuals of the mating pool, the new GA population, $P(t + 1)$, is obtained.

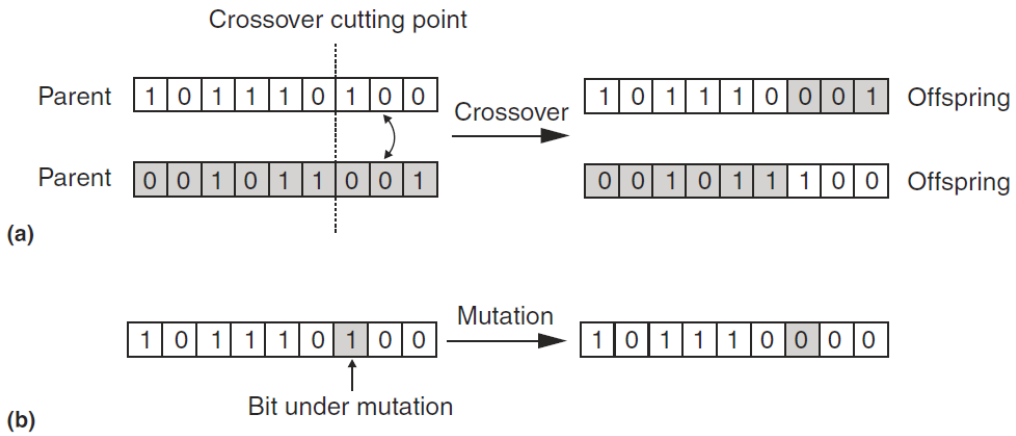


Fig.III.12. (a) Example of single-point crossover for a binary genetic algorithm. A cutting point is chosen randomly and corresponding portions of parents recombined in order to obtain two offspring. (b) Example of mutation for a binary genetic algorithm. A bit is randomly selected and its allele value changed.

III.4.3.1.5. Variants of the Holland's Model

The model proposed by Holland has inspired first theoretical studies and applications of GAs. However, it is not always natural or convenient to use bit string encoding (not always is proportional fitness of the best selection method) and Holland's genetic operators are not always the most effective and appropriate [15]. Furthermore, it is not a good idea to lose the best individuals that are found during the GA evolution, because it has been shown that elitism improves performances in both single-objective GAs and MOGAs [16]. For these reasons, new models have been proposed from the late 1980s, which differ from the original Holland's model in the genotype-encoding scheme, in the adopted genetic operators and selection strategy.

III.4.3.1.6. Selection methods and elitism

Selection is one of the fundamental processes of a GA because it eliminates individuals with lower fitness and creates one or more copies of individuals with higher fitness from which individuals of the new population are generated. The selection operator has a substantial effect on the dynamics of GAs: too much selective pressure may result in an overly rapid convergence, by entrapping the algorithm in a local optimum from which it will be unable to exit; on the other hand, weak selective pressure can lead to an excessive increase in the amount of time required to find an acceptable solution.

Selection operators can replace the entire population (in this case, we talk of generational GAs) or only part of it (generation gap GAs). Furthermore, a steady-state GA is obtained if at most two individuals are replaced. In addition, the operator can select an individual once or more than once. The first case refers to a selection operator without replacement, in the sense that the selected individual is not reinserted back into the old population after mating and, therefore, cannot be selected again. To the contrary, in the second case, the chosen individual is reinserted in the old population and can, there-fore, be selected again, by producing more offspring.

Both in steady-state, generation gap and generational GAs, it may happen that the best individuals are lost in the transition to the subsequent generation. The models that ensure the survival of best individuals are called elitist (or k-elitist, where k is the number of the best individuals that are pre-served and copied in the new population).

Besides the proportional selection operator proposed by Holland, the tournament selection is one of the most used in practical applications. The latter, as well as other selection operators (e.g., the Boltzmann and the rank-based ones), was introduced in order to have less-selective pressure with respect to the proportional one [15]. In the most common type of tournament selection, two individuals are chosen at random from the current population and a number $c \in [0, 1]$ is randomly generated. If c is less than a prefixed parameter $r \in [0, 1]$, for example $r = 0.75$, the most fit individual wins the tournament and is selected, otherwise the less fit is the winner. In addition, if the scheme with replacement is applied, the two individuals are reintegrated in the old population and may be selected again.

III.4.3.2. NSGA-II

The non-dominated sorting GA or NSGA-II procedure [17] for finding multiple Pareto-optimal solutions in a MOP has the following three features:

1. It uses an elitist principle,
2. It uses an explicit diversity preserving mechanism, and
3. It emphasizes the non-dominated solutions.

In NSGA-II, the offspring population Q_t is first created by using the parent population Pt and the usual genetic operators [18]. Thereafter, the two populations are combined to form R_t of size $2N$. Then, a non-dominated sorting is used to classify the entire population R_t. Once the non-dominated sorting is over, the new population is filled by solutions of different non-dominated fronts, one at a time. The filling starts with the best non-dominated front and continues with solutions of the second non-dominated front, followed by the third non-dominated front, and so on. Since the overall population size of R_t is $2N$, not all fronts may be accommodated in N slots available in the new population. All fronts which could not be accommodated are simply deleted. When the last allowed front is being considered, there may exist more solutions in the last front than the remaining slots in the new population. This scenario is illustrated in Figure III.13. Instead of arbitrarily discarding some members from the last acceptable front, the solutions which will make the diversity of the selected solutions the highest are chosen. The NSGA-II procedure is outlined in the following.

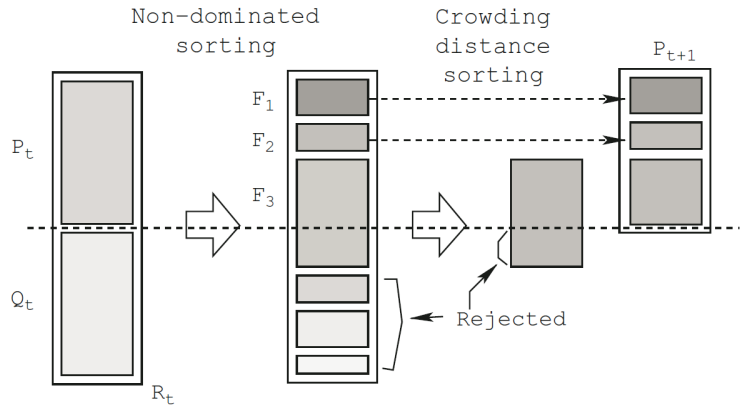


Fig.III.13. Schematic of the NSGA-II procedure.

NSGA-II

Step 1: Combine parent and offspring populations and create $R_t = P_t \cup Q_t$. Perform a non-dominated sorting to R_t and identify different fronts: F_i, $i = 1, 2, ...$, etc.

Step 2: Set new population $P_{t+1} = \emptyset$. Set a counter $i = 1$.
Until $|P_{t+1}| + |F_i| < N$, perform $P_{t+1} = P_{t+1} \cup F_i$ and $i = i + 1$.

Step 3: Perform the Crowding-sort (F_i, $< c$) procedure and include the most widely spread ($N - |P_{t+1}|$) solutions by using the crowding distance values in the sorted F_i to P_{t+1}.

Step 4: Create offspring population Q_{t+1} from Pt+1 by using the crowded tournament selection, crossover and mutation operators.

In Step 3, the crowding-sorting of the solutions of front i (the last front which could not be accommodated fully) is performed by using a crowding distance metric, which we describe later. The population is arranged in descending order of magnitude of the crowding distance values. In Step 4, a crowding tournament selection operator, which also uses the crowding distance, is used.

The crowded comparison operator ($< c$) compares two solutions and returns the winner of the tournament. It assumes that every solution i has two attributes:

1. A non-domination rank r_i in the population,
2. A local (d_i) in the population.

The crowding distance d_i of a solution i is a measure of the normalized search space around i which is not occupied by any other solution in the population. Based on these two attributes, we can define the crowded tournament selection operator as follows.

Definition III.7: Crowded tournament selection operator. A solution i wins a tournament with another solution j if any of the following conditions are true:

1. If solution i has a better rank, that is, $r_i < r_j$.
2. If they have the same rank but solution i has a better crowding distance than solution j, that is, $r_i = r_j$ and $d_i > d_j$.

The first condition makes sure that the chosen solution lies on a better non-dominated front. The second condition resolves the tie of both solutions being on the same non-dominated front by deciding on their crowded distance. The one residing in a less crowded area (with a larger crowding distance d_i) wins. The crowding distance di can be computed in various ways. However, in NSGA-II, we use a crowding distance metric, which requires O (MN log N) computations.

To get an estimate of the density of solutions surrounding a particular solution i in the population, we take the average distance of two solutions on either side of solution i along each of the objectives. This quantity di serves as an estimate of the perimeter of the cuboid formed by using the nearest neighbors as the vertices (we call this the crowding distance). In Figure III.14, the crowding distance of the ith solution in its front (marked with filled circles) is the average side-length of the cuboid (shown by a dashed box). The following algorithm is used to calculate the crowding distance of each point in the set F.

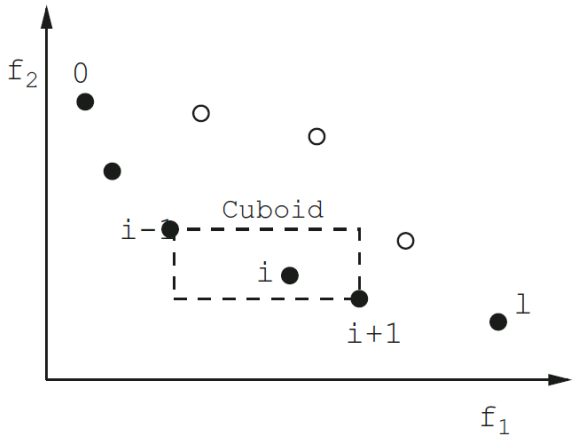


Fig.III.14. The crowding distance calculation.

Crowding distance assignment procedure: Crowding-sort (F, $< c$)

Step C1: Call the number of solutions in F as $l = |F|$. For each i in the set, first assign $d_i = 0$.

Step C2: For each objective function $m = 1, 2, ..., M$, sort the set in worse order of f_m or, find the sorted indices vector: $I^m = \text{sort}(f_m, >)$.

Step C3: For $m = 1, 2, ..., M$, assign a large distance to the boundary solutions, or $d_{I_1^m} = d_{I_l^m} = \infty$, and for all other solutions $j = 2$ to $(l - 1)$, assign

$$d_{I_j^m} = d_{I_j^m} + \frac{f_m^{(I_{j+1}^m)} - f_m^{(I_{j-1}^m)}}{f_m^{\max} - f_m^{\min}}.$$

Index I_j denotes the solution index of the jth member in the sorted list. Thus, for any objective, I_1 and I_l denote the lowest and highest objective function values, respectively. The second term on the right-hand side of the last equation is the difference in objective function values between two neighboring solutions on either side of solution I_j. Thus, this metric denotes half of the perimeter of the enclosing cuboid with the nearest-neighboring solutions placed on the vertices of the cuboid (Fig.III.14). It is interesting to note that for any solution i the same two solutions $(i+1)$ and $(i-1)$ need not be neighbors in all objectives, particularly for $M \geq 3$. The parameters $f_m^{max}$and f_m^{min}can be set as the population-maximum and population-minimum values of the mth objective function. The above metric requires M sorting calculations in Step C2, each requiring O (MN log N) computations. Step C3 requires N computations. Thus, the complexity of the above distance metric computation is O (MN log N) and the overall complexity of one generation of NSGA-II is O (MN^2), governed by the non-dominated sorting procedure.

III.4.4. Sample Simulation Results

In this section, we show the simulation results of NSGA-II on (SCH1) test problems which is simple two-objective problem with a convex Pareto-optimal front:

$$\text{SCH1}: \begin{cases} \text{Minimize } f_1(x) = x^2 \\ \text{Minimize } f_2(x) = (x-2)^2 \\ -10^3 \leq x \leq 10^3. \end{cases} \qquad \text{(III.5)}$$

NSGA-II is run with a population size of 100 and for 250 generations. Figure III.15 shows that NSGA-II converges on the Pareto-optimal front and maintains a good spread of solutions. In comparison to NSGA-II, another competing EMO method—the Pareto archived evolution

strategy (PAES) [19]—is run for an identical overall number of function evaluations and an inferior distribution of solutions on the Pareto-optimal front is observed.

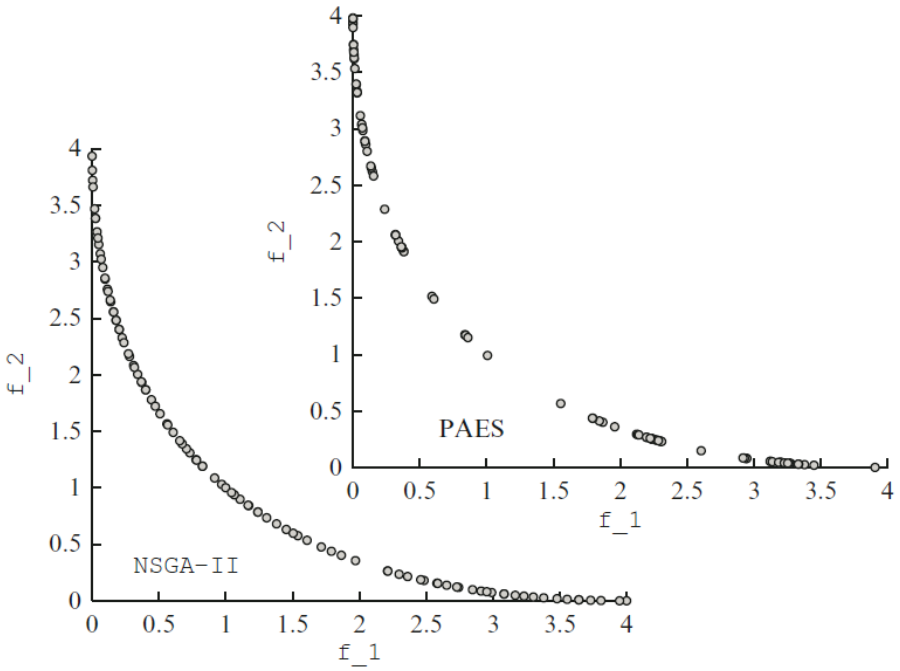


Fig.III.15. NSGA-II finds better spread of solutions than PAES on SCH.

III.5. Conclusion

For the past two decades, the usual practice of treating MOPs by scalarizing them into a single objective and optimizing it has been seriously questioned. The presence of multiple objectives results in a number of Pareto-optimal solutions, instead of a single optimum solution. In this chapter, we have discussed the use of an ideal multi-objective optimization procedure which attempts to find a well-distributed set of Pareto-optimal solutions first. It has been argued that choosing a particular solution as a post-optimal event is a more convenient and pragmatic approach than finding an optimal solution for a particular weighted function of the objectives. Be-sides introducing the multi-objective optimization concepts, this chapter has also presented MGA which is the most commonly used MOEAs. Besides finding the multiple Pareto-optimal solutions, the suggested ideal multi-objective optimization procedure has another unique advantage. Once a set of Pareto-optimal solutions are found, they can be analyzed. The principle behind the transition from the optimum of one objective to that of other objectives can be investigated as a post-optimality analysis. Since all such solutions are optimum with respect to certain trade-off between objectives, the transition should reveal interesting knowledge on an optimal process of sacrifice of one objective to get a gain in other objectives.

References

[1] Deb K (2001) Multi-objective optimization using evolutionary algorithms. Wiley, Chichester.

[2] Coello CAC, Lechuga MS (2002) MOPSO: a proposal for multiple objective particle swarm optimization. In: Proceedings of the CEC 2002, vol 2. IEEE, Piscataway, Honolulu, USA, pp. 1051–1056.

[3] Goh CK, Tan KC (2009) Evolutionary multi-objective optimization in uncertain environments: issues and algorithms. Springer, Berlin.

[4] Fonseca C, Fleming P, Zitzler E, Deb K, Thiele L (eds) (2003) Proceedings of the EMO-2003, Faro. LNCS 2632. Springer, Heidelberg.

[5] Ehrgott M, Fonseca CM, Gandibleux X, Hao JK, Sevaux M (eds) (2009) Proceedings of the EMO-2009, Nantes. LNCS 5467. Springer, Heidelberg.

[6] Coello CAC (2003) http://www.lania.mx/~ccoello/EMOO/

[7] Deb K (2003) Unveiling innovative design principles by means of multiple conflicting objectives. Eng Optim 35:445–470.

[8] Cormen TH, Leiserson CE, Rivest RL (1990) Introduction to algorithms. Prentice- Hall, New Delhi.

[9] Chankong V, Haimes YY (1983) Multiobjective decision making theory and methodology. North-Holland, New York.

[10] Kung HT, Luccio F, Preparata FP (1975) On finding the maxima of a set of vectors. J Assoc Comput Mach 22:469–476.

[11] Miettinen K (1999) Nonlinear multiobjective optimization. Kluwer, Boston Mostaghim S, Teich J (2003) Strategies for finding good local guides in multi-objective particle swarm optimization (MOPSO). In: Proceedings of the 2003 IEEE symposium on swarm intelligence, Indianapolis. IEEE, Piscataway, pp 26–33.

[12] Haimes YY, Lasdon LS, Wismer DA (1971) On a bicriterion formulation of the problems of integrated system identification and system optimization. IEEE Trans Syst Man Cybern 1:296–297.

[13] Holland, J.H., 1967. Nonlinear environments permitting efficient adaptation. Proceedings of Computer and Information Sciences – II. Academic Press, New York, NY, pp. 147–164.

[14] Holland, J.H., 1975. Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor, 228 pp.

[15] Mitchell, M., 1996. An Introduction to Genetic Algorithms. MIT Press, Cambridge, 217 pp.

[16] Rudolph, G., Agapie, A., 2000. Convergence Properties of Some Multi-Objective Evolutionary Algorithms. Proceedings of the 2000 Conference on Evolutionary Computation, La Jolla, CA, USA, 16–19 July, 2000, pp. 1010–1016.

[17] Deb K, Jain S (2002) Running performance metrics for evolutionary multi-objective optimization. In: Proceedings of the 4th Asia-Pacific conference on simulated evolution and learning (SEAL-02), Singapore, pp 13–20.

[18] Goldberg DE (1989) Genetic algorithms for search, optimization, and machine learning. Addison-Wesley, Reading.

[19] Knowles JD, Corne DW (2000) Approximating the non-dominated front using the Pareto archived evolution strategy. Evol Comput J 8:149–172.

Chapter IV:

Multiobjective optimization for IMRT

Chapter IV: Multiobjective optimization for IMRT

IV.1. Introduction

In the inverse planning of a radiation treatment plan, a collection of parameters (beams and fluences) is algorithmically computed for a predefined treatment plan in order to satisfy the prescribed doses and constraints. Inverse treatment planning enables the modeling of exceedingly complex treatment planning problems, and optimization is crucial to the procedure's success. Intensity modulated radiation therapy (IMRT) is a sort of inverse treatment planning in which the radiation beam is modulated by a multileaf collimator, as shown in Figures IV.1 and IV.2.

In Figure IV.1(a), Multileaf collimators (MLC), allow the beam to be transformed into a grid of smaller beamlets with different intensities as depicted in figure IV.1(b). Beamlets do not exist physically, notwithstanding the depiction in Figure IV.1(b). The MLC leaves movement as shown in figure IV. 1(a), which block part of the beam during parts of the delivery time, generate their existence. Both sides of the MLC include adjustable leaves that can be positioned at any beamlet grid boundary. MLC has two modes of operation: dynamic collimation and multiple static collimations. The leaves in the first situation move continuously during irradiation contrary to the second situation" step and shoot mode", where the leaves are programmed to open a desired aperture during each segment of the delivery, and radiation is on for a given fluence time or intensity. This approach provides a distinct set of intensity maps (the set of chosen beam angles) as shown in Figure IV. 1. (b). In this case, we'll consider multiple static collimations.

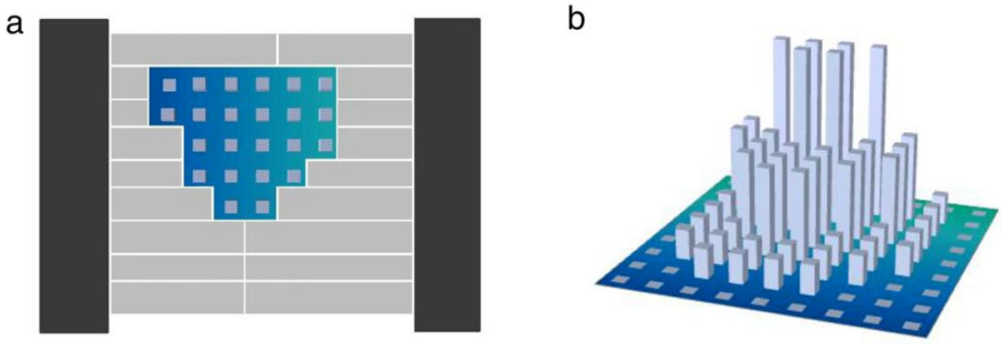


Fig.IV.1. A multileaf collimator (with nine pairs of leaves) (a), and a beamlet intensity map (9 × 9) (b) illustrations.

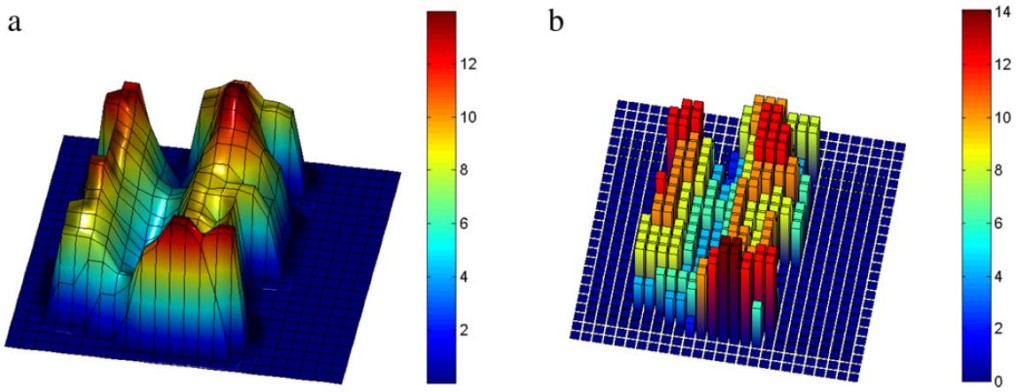


Fig.IV.2. A beam's Ideal theoretic (a) and deliverable (b) fluence.

A beamlet-based approach is a common strategy to solve inverse planning in IMRT optimization challenges. The volume of each structure is discretized in voxels for optimization purposes (volume elements). A three-dimensional coordinate is assigned to each voxel in a structure (*x, y, z*). Consider the case where there is $m \times n$ beamlets recognized by the index pair (p, q). w (θ, p, q) is the weight (intensity) of the beamlet (p, q) delivered over an angle θ. The total dose, D (x, y, z), that a voxel (x, y, z) gets is calculated using the superposition principle as follows:

$$D(x,y,z) = \sum_{(\theta,p,q)} w(\theta,p,q) \cdot d_{(\theta,p,q)}(x,y;z)$$

where $d_{(\theta,p,q)}$(x, y, z) is the dose delivered to voxel (x, y, z) by beamlet (p, q) from angle θ. As shown in figure IV. 3.

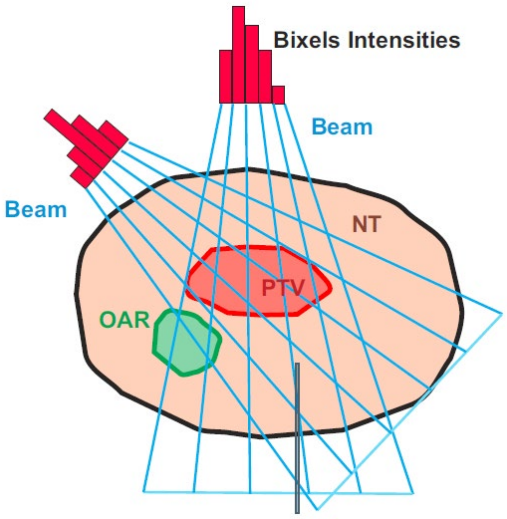


Fig.IV.3. IMRT dose optimization Principle. The planning target volume (PTV), one organ at risque (OAR) and the contours of the body are depicted. The challenge is to determine the intensities of the tiny subdivisions (bixels) of each beam, in order to get the best dose distribution possible.

This beamlet-based method generates a large-scale programming issue with thousands of variables (beamlets) and hundreds of thousands of constraints (dose–volume). The clinical treatment effect is determined by the plan's quality, which is conditional to the programming models and resolution methods. As a result of the overall optimization problem complexity, the treatment planning is generally separated into three smaller problems: The geometry problem, intensity problem, and realization problem which can be solved independently.

The geometry problem entails utilizing optimization methods to identify the smallest number of beams and corresponding orientations that satisfy the treatment goals [1–3]. In practical practice, the number of beams is usually considered to be determined a priori by the treatment planner, and the beam orientations are still decided manually by the treatment planner, who mostly depends on his or her experience. After identifying which beam angles should be employed, the intensity (or fluence map) problem, which is the problem of determining the appropriate beamlet weights for the fixed beam angles, is solved. As a result, the relation between the geometry problem and the intensity problem is clear, because the geometry problem's angle values are an input to the intensity problem (despite whether the treatment planner computes or manually selects them).

For the intensity problem, a variety of mathematical optimization models and techniques have been developed, including linear models [4,5], mixed integer linear models [6,7], nonlinear models [8,9], and multiobjective models [10,11]. The intensity problem yields a set of ‘‘continuous" fluence maps that are optimized (one for each beam angle).

The intensity profiles, in current inverse planning systems for IMRT are determined through a computerized optimization process based on specified dose prescriptions for targets and organs at risk, whereby so-called weight factors must be allocated to each structure. The optimization outcome is simply the treatment plan with the best number, and the quality of a plan is judged by a single number calculated by adding together the deviations from the prescriptions.

The problem is that the resulting compromise between the opposing planning goals is frequently not clinically acceptable, necessitating many optimization runs with varied parameters until an acceptable compromise is established [12]. This trial and error process can take a long time, and even once a plan is accepted for treatment, it's unclear whether a better plan for the patient would have resulted if the planner had attempted a few more parameter settings. All of these issues combine to mean that the full promise of IMRT is not realized for some patients due to constraints in the inverse planning process.

Multiobjective (also known as multicriteria) optimization (MO) is a potential strategy for overcoming the current situation. The optimization outcome is no longer a single plan, but rather a database of plans, each of which represents a so-called Pareto-optimal solution [13] that can't be improved in one criterion without worsening in at least one other. The planner can experience the sensitivity to changes in certain structures by interactively browsing the database and deciding on the clinically optimal compromise.

In the context of radiation therapy, the Pareto concept and MO have been used to optimize beam angles [14], brachytherapy [15], radiosurgery [16], and external radiotherapy [17-20], mostly in a research setting.

To obtain a multiobjective optimization problem's Pareto solution set, many multiobjective evolutionary algorithms including Multiobjective Genetic Algorithm (MGA) by using its multiobjective Non-dominated Sorting Genetic Algorithm (NSGA-II), have been proposed recently. They've been employed in a variety of fields due to their excellent efficiency, however they're rarely used in inverse planning for multiobjective optimization.

In this work, multiobjective optimization of IMRT planning was studied based on inverse planning research by the Computaional Envierenement Radioherapy Research System (named CERR) [21]. The mathematical modeling was presented first, in which a multiobjective optimization problem with various constraints was created from the clinical needs for a treatment plan. The NSGA-II was then used to improve the model. Finally, a clinical example was put to the test. The findings of the pareto front reveal that the non-dominated solutions obtained were distributed equally. The associated dose distribution of one of the non-dominated solution set's solutions thus approached the expected dose distribution while also meeting the dose-volume limitations. The clinical requirements were better satisfied, and the planner was able to choose the best treatment plan from the non-dominated solution set. With the method we offer, the planner will no longer need to go through a trial and error procedure to find the best plan, resulting in a significant increase in efficiency.

IV.2. Method

IV.2.1 Description of the IMRT optimization problem

The goal of mathematical modeling is to determine the optimization objective function, which measures the efficacy of a chosen plan, and its selection is critical for radiotherapy treatment planning optimization. The 'physical' objective function and the 'biological' objective function are the two sorts of objective functions. Due to its widespread use in the commercial Treatment Planning System, the 'physical' objective function, which provides a link between the output dose distribution and the input beam parameters, is employed in our situation (TPS). Here we describe the fluence-based IMRT optimization problem as follows:

The dose influence matrix *Dij* is the main entity used for optimization. It contains the dose delivered to each voxel *i* per unit intensity of beamlet *j*.

The dose to voxel *i* is given by:

$$d_i = \sum_j D_{ij} x_j \qquad \text{(IV.1)}$$

where x_j is the fluence value of the *j*th beamlet.

We assume that the *D* matrix represents the whole set of beamlets from all the beams for a given set of beams. In other words, below *D* is read as a concatenation of each beam's distinct *D* matrices. This notation simplifies the problem by allowing us to loop over all beamlets rather than looping over the beams. Let *d* be the voxel doses vector and *x* be the beamlet fluences vector. The linear link between the beamlet vector and the dosage distribution given in Equation is the most important mapping (IV.1). The following is a generic formulation of the IMRT optimization issue written in the form of a matrix vector product $Dx = d$:

$$\text{minimize } f(d) \text{ such that} \begin{cases} Dx = d \\ d \in C \\ x \geq 0 \end{cases} \qquad \text{(IV.2)}$$

Choosing the mean dose to a crucial structure as $f(d)$ and using the constraint set *C* to trigger upper bounds for all voxels and supplementary lower bounds for the target voxels would result in a linear program.

The linear mapping from the fluence values x to the voxel doses d, as described in Equation (IV.2), is used in the optimization formulation (IV.1). As a result, the problem is a convex optimization problem because the function $f(d)$ is convex and the constraint set C is convex.

IV.2.2. NSGA-II optimization algorithm

Because they are characterized by a population of solution candidates and can produce a set of approximate solutions in a simulated run, evolutionary algorithms are popular for solving multiobjective optimization problems. Because of its validity, the NSGA-II has been used to numerous domains as a representative of multi-objective evolutionary algorithms [9]. As a result, in this work, the NSGA-II was introduced for inverse planning. The following were the procedures:

1. Initialization: the population size (N) and maximum evolutionary generation (EG) were determined, and then N individuals from a parent population (P_0) were generated at random;
2. By decoding the population P_0, N groups of field parameters were obtained, and then the point dose values in the target and critical structure were determined to compute the objective functions and constraint values using Equations (IV.1) and (IV.2);
3. The objective functions and constraint functions were used to compute the non-dominated rank and individuals crowding distance in population P_0 [9, 10];
4. The binary tournament selection was used to choose N individuals into the next population Q_0, taking into account the individual's non-dominated rank and the crowding distance;
5. Individuals in Q_0 carried out an evolutionary process (which included crossover and mutation) and then repeated steps 2 and 3 for Q_0; after completed, the process moved on to step 6.
6. The offspring population (Q_0) and the parent population (P_0) were united into a population Q of $2N$ individuals, and N individuals of the following generation's parent population were created via tournament selection from the Q individuals.
7. The optimization process would be ended and the obtained Pareto solutions (field parameters) exported if the iteration times or other conditions were fulfilled; otherwise, the process would proceed to step 2.

IV.3. Results and discussion

IV.3.1. Test case

A patient with a clinical liver tumor with 168 CTs slices was chosen to test the method. the DICOM CT data was imported into CERR and the CT scan was then resampled to the voxel sizes shown in Table 1 [22].

	Liver
Number of beam angles	07
Noncoplanar	yes
Beamlet size [cm]	1 × 1
Voxel resolution (LR,AP,SI) [mr	(3.0, 3.0, 2.5)
Voxel grid size (LR,AP,SI)	(217, 217,168)
Number of target voxels	6954
Number of voxels in patient	1,927,357

Table IV.1. Summary of patient caracteristics

The PTV and OAR including heart, liver and entrance were contoured as shown in figure IV.4.

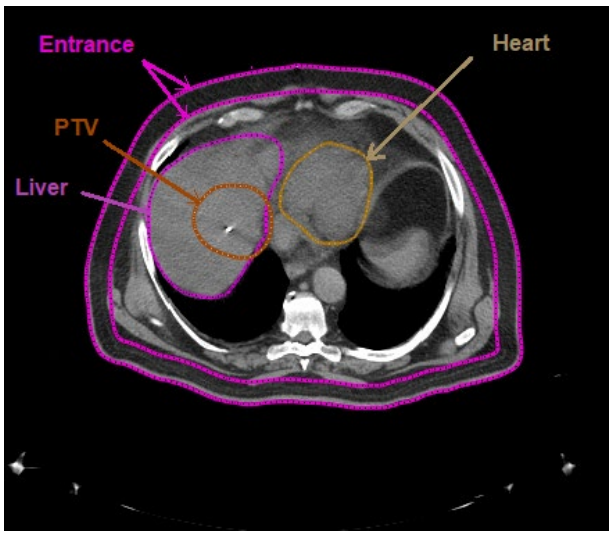


Fig.IV.4. Axial view for PTV and OARs Contour.

Seven non-coplanar beams with different orientations were adopted: at (gantry, couch) angles (58°, 0°), (106°, 0°), (212°, 0°), (328°, 0°), (216°, 32°), (226°, -13°), (296°, 17°). (Coplanar refers to beams where the couch angle is fixed at 0°).

The dose influence matrices for this case were created with CERR. Which use the quadrant infinite beam (QIB) model [23,24] is a pencil beam type dose calculation algorithm. This approach computes *Dij* quickly by using precalculated integration values.

Gray per monitor unit (Gy/MU) is the dose influence matrix unit. The beamlet intensity unit (MU) is defined as 100 MU is a dose of 1 Gy delivered in 10 cm depth in water in the middle of a 10 cm × 10 cm radiation field. The dose-influence matrix is measured in Gy/MU in order to allow research involving treatment delivery time and/or changing dose rates.

Regarding the specifics of the dose computation, we used the default values in the CERR IMRT GUI (Gaussian primary and scatter radiation, exponential scatter method, 6 Megaelectron-volts beams).

IV.3.2. Algorithm parameters

To solve the liver case, we utilized the Matlab NSGA-II MGA solver (gamultiobj), which minimizes the weighted sum of the mean doses to the liver, heart, and normal tissue in the entrance region, subject to the restrictions that every PTV voxel receives a dose greater than one, as shown in table IV.2.

Objective	min (mean Liver + mean Heart+ 0.6*mean entrance)
Constraints	PTV >= 1
	x <= 25

Table IV.2. MGA formulation for the liver case

The algorithm's parameters were set as follows: The population size was 150; the maximum generation was 1000; the crossover probability was 0.6; the mutation probability was 0.01; the variables were handled as binary code; the binary bits of variables differed in accuracy.

Because the test case had two objectives that were contradicting, there is no one best solution, but rather a range of feasible solutions of comparable quality (Pareto solutions). Using the optimization, 27 Pareto solutions were found. Figure IV.5 depicts the distributions of the two objective values of 36 solutions.

Figure IV.6 shows an isodose and colorwash dose distributions for the first solution.

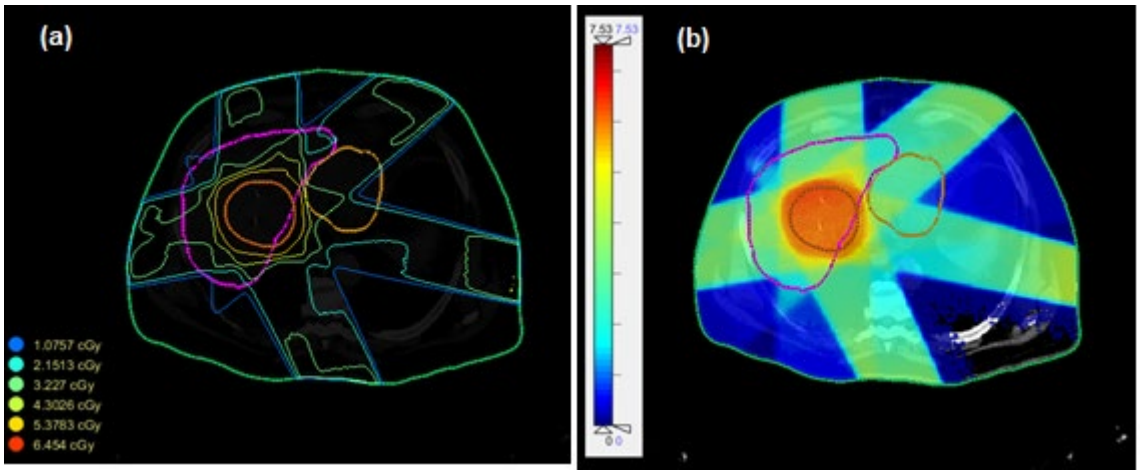


Fig.IV.6. shows an isodose (a) and colorwash (b) dose distributions for the first solution.

Figure IV.7 shows the equivalent Dose Volume Histograms (DVH) of the PTV, OARs: Liver, Heart, and Entrance.

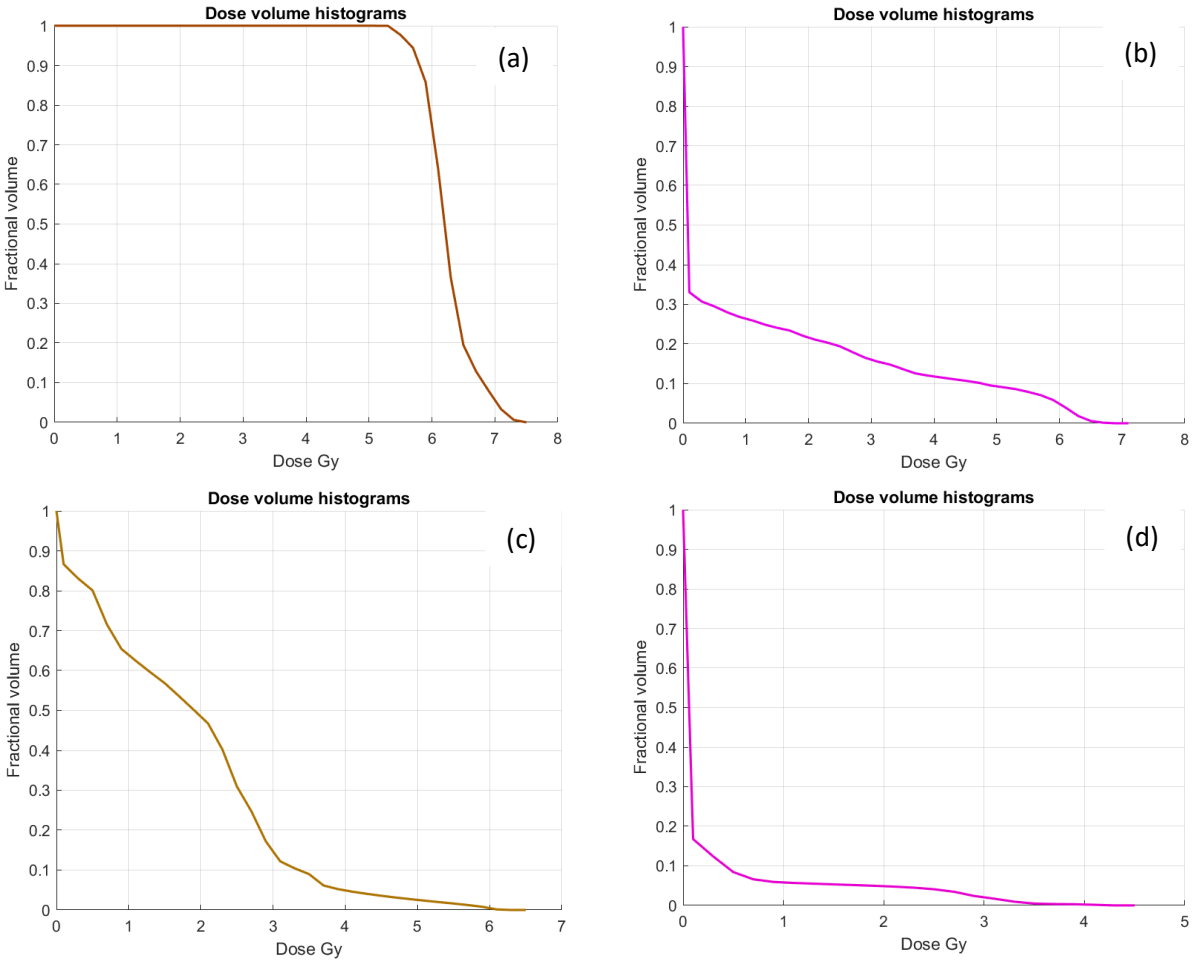


Fig.IV.7. Dose Volume Histograms (DVH) of the (a) PTV and OARs: (b) Liver, (c) Heart and (d) Entrance.

IV.4. Conclusion

In this chapter, IMRT fluence map optimization was mathematically modeled as a multiobjective optimization problem, taking advantage of both the objective function based on the dose distribution and the objective function based on the dose-volume constraints, and then a multiobjective evolutionary algorithm MGA based on the NSGA-II was introduced to solve the problem. Clinical test results revealed that numerous optimal solutions might be found, giving planners the greatest option for balancing different objectives and dose-volume constraints.

The proposed method in this chapter provides a Pareto optimal solution set for planners to choose from, rather than forcing an inexperienced user to import weighting components iteratively. As a result, this technique is more precise and adaptable to meet practical clinical needs.

References

[1] H. Rocha, J.M. Dias, B.C. Ferreira, M.C. Lopes, Direct search applied to beam angle optimization in radiotherapy design, Inescc Research Report 06/2010, ISSN: 1645–2631. Available at: www.inescc.pt/documentos/6_2010.PDF.

[2] D. Craft, Local beam angle optimization with linear programming and gradient search, Phys. Med. Biol. 52 (2007) 127–135.

[3] M. Ehrgott, A. Holder, J. Reese, Beam selection in radiotherapy design, Linear Algebra Appl. 428 (2008) 1272–1312.

[4] H.E. Romeijn, R.K. Ahuja, J.F. Dempsey, A. Kumar, A column generation approach to radiation therapy treatment planning using aperture modulation, SIAM J. Optim. 15 (2005) 838–862.

[5] H.E. Romeijn, R.K. Ahuja, J.F. Dempsey, A. Kumar, J. Li, A novel linear programming approach to fluence map optimization for intensity modulated radiation therapy treatment planing, Phys. Med. Biol. 48 (2003) 3521–3542.

[6] E.K. Lee, T. Fox, I. Crocker, Integer programing applied to intensity-modulated radiation therapy treatment planning, Ann. Oper. Res. 119 (2003) 165–181.

[7] F. Preciado-Walters, M.P. Langer, R.L. Rardin, V. Thai, Column generation for IMRT cancer therapy optimization with implementable segments, Ann. Oper. Res. 148 (2006) 65–79.

[8] D.M. Aleman, D. Glaser, H.E. Romeijn, J.F. Dempsey, Interior point algorithms: garanteed optimality for fluence map optimization in IMRT, Phys. Med. Biol. 55 (2010) 5467–5482.

[9] S. Spirou, C.-S. Chui, A gradient inverse planning algoritm with dose–volume constraints, Med. Phys. 25 (1998) 321–333.

[10] H.E. Romeijn, J.F. Dempsey, J. Li, A unifying framework for multi-criteria fluence map optimization models, Phys. Med. Biol. 49 (2004) 1991–2013.

[11] C. Thieke, K.H. Kufer, M. Monz, A. Scherrer, F. Alonso, U. Oelfke, P.E. Huber, J. Debus, T. Bortfeld, A new concept for interactive radiotherapy planning with multicriteria optimization: first clinical evaluation, Radiother. Oncol. 85 (2007) 292–298.

[12] Hunt MA, Hsiung CY, Spirou SV, et al. Evaluation of concave dose distributions created using an inverse planning system. Int J Radiat Oncol Biol Phys 2002;54:953–62.

[13] Pareto V. Manual of political economy. New York, New York: A. M. Kelley; 1971 [Translation of Manuale di economia politica 1906].

[14] Haas OCL, Burnham KJ, Mills JA. Optimization of beam orientation in radiotherapy using planar geometry. Phys Med Biol 1998;43:2179–93.

[15] Lahanas M, Baltas D, Zamboglou N. Anatomy-based three-dimensional dose optimization in brachytherapy using multi-objective genetic algorithms. Med Phys 1999;26:1904–18.

[16] Yu Y, Zhang JB, Cheng G, Schell MC, Okunieff P. Multi-objective optimization in radiotherapy: applications to ste-reotactic radiosurgery and prostate brachytherapy. Artif Intell Med 2000;19:39–51.

[17] Cotrutz C, Lahanas M, Kappas K, Baltas D. A multiobjective gradient based dose optimization algorithm for external beam conformal radiotherapy. Phys Med Biol 2001;46:2161–75.

[18] Craft DL, Halabi TF, Shih HA, Bortfeld TR. Approximating convex Pareto surfaces in multiobjective radiotherapy plan-ning. Med Phys 2006;33:3399–407.

[19] Hoffmann AL, Siem AY, den Hertog D, Kaanders JH, Huizenga H. Derivative-free generation and interpolation of convex Pareto optimal IMRT plans. Phys Med Biol 2006;51:6349–69.

[20] Schreibmann E, Lahanas M, Xing L, Baltas D. Multiobjective evolutionary optimization of number of beams, their orienta-tion and weights for IMRT. Phys Med Biol 2004;49:747–70.

[21] Deasy J, Blanco A, Clark V: CERR: A computational environment for radiotherapy research. Med Phys 2003, 30(5):979–985.

[22] Craft et al.: Shared data for intensity modulated radiation therapy (IMRT) optimization research: the CORT dataset. GigaScience 2014 3:37.

[23] Ahnesjö A, Saxner M, Trepp A: A pencil beam model for photon dose calculation. Med Phys 1992, 19(2):263–273.

[24] Kalinin E, Deasy J: A method for fast 3-D IMRT dose calculations: The quadrant infinite beam (QIB) algorithm. Med Phys 2003, 30(6):1348–1349.

Chapter V:

Double Graphene-Gate Junctionless Radiation Sensitive FET (DGG JL RADFET) Dosimeter

Chapter V: Double Graphene-Gate Junctionless Radiation Sensitive FET (DGG JL RADFET) Dosimeter

V.1. Introduction

Compared to existing double gate MOSFETs, the Junction-less Double Gate Field Effect Transistor (JL DG FET) has proved its advantages as an outstanding structure to offer good control over the channel and better immunity against short-channel phenomena **[1].** Taking this impressive property of JL DG FET into account, studies explored the Radiation Sensitive Field Effect Transistor (RADFET) dosimeter to quantify the captivated dosage as a result of the field impact produced by the localized charges **[2, 3].** RADFETs can be employed for radiation-prone nuclear, space and radiotherapy applications thanks to their low energy utilization, efficiency and compliance with normal CMOS technologies **[4-6].** In **[7],** a thorough analysis of the radiation sensor and dosimeter was described. The basic RADFET dosimeter principle relies on the calculation of the threshold voltage change followed by the conversion of such difference to the absorbed dosage. Due to their advantages over conventional dosimetry systems, different MOSFET based dosimeters have been produced in recent decades **[8-11].** Instantaneous and non-destructive measurements, basic calibration, high sensitivity, low energy utilization, reliability and large dosage interval are the key benefits of MOS-based dosimeters over standard dosimeters. The pMOS dosimeter displays fading with thicker gate oxide **[12, 13],** amidst certain benefits than most dosimeters. However, as the length of the interface gate reaches the nanoscale level, the output of the device is drastically influenced by numerous short channel effects **[14].** Several relevant contributions have been reported to boost the sensitivity of RADFET through considering gate stack pMOS characterized by two layers of gate oxide Dual Dielectric materials. It is worthy to mention that pMOS Dosimeter, GAA MOSFET and JL DG FET are actually recommended for radiation sensor owing to their processing benefits and high immunity to short-channel effects **[15-18]**. Because of the promising properties of Graphene, some experiments have focused on the control of graphene work function by imposing an electric field or chemical doping, which yields values in the range 3.5 to 5.16 eV **[19, 20].** The graphene work function can be also adjusted by modifying the number of graphene layers **[21].** Herein, an analytical modeling framework involving the regional method for JL DG RADFET and graphene as gate material is presented to predict the variance of surface potential and threshold voltage reflected by localized charges related to irradiation-induced damage. Furthermore, we assess the influence of adjusting monolayer graphene work function on the response of JL DG RADFET. As a first step, the

proposed device is considered without optimization and later such aspect is carried out by taking into account the impact of localized charges induced by radiation, Si body width, and graphene work function. To the best of our knowledge, only a few design policies based on junctionless dosimeter aspect using graphene as gate material, and device global optimization including degradation effects are reported to enhance the structure response for dosimeter applications. Therefore, and in order to obtain a more accurate description of the efficiency of JL-DGG RADFET under damage conditions caused by radiation effects, the analytical model is advantageous not just for compact modeling, but also for radiation degradation modeling, which is a very interesting issue, particularly in the understanding of the device's sensitivity behavior. In addition, it would be necessary to build fitness functions for the global optimization of system reliability and performance on the basis of established analytical models. It should be mentioned that a wide range of stochastic tools have been investigated for the tuning of configuration parameters for semiconductor based devices [22, 23]. However, comparison between different metaheuristics may be an intractable task according to the famous No Free Lunch Theorem, which states that no optimization algorithm can outperform any other under any metric over all problems [24]. This why we have focused only on the investigation of MGA in order to get a more accurate view regarding the performance of the proposed design. The sensitivity of the optimized JL-DGG RADFET is compared to conventional RADFETs in order to assess the benefit of using the suggested strategy combining analytical modeling and multi objective optimization method. The findings reveal that in contrast with other RADFETs, the suggested strategy demonstrates its efficacy in improving both sensitivity and electrical efficiencies.

The remain of this chapter is organized as follows. Section 2 is dedicated to the description of the proposed RADFET device. In Section 3, we highlight the basic steps of the elaboration of the associated analytical modeling framework. The theoretical background of MGA is showcased in Section 4 including different genetic operators and offsprings encoding. The obtained simulation results are presented and interpreted in Section 5. Finally, some conclusions and future work directions are provided in Section 6.

V.2. Device architecture

Figure V.1 displays a schematic design of the Junction-less Double graphene-Gate RADFET Dosimeter, where t_{ox} and t_b represent the oxide and the channel thicknesses, respectively. L_d and

L_s represent the damaged area length and the safe area, respectively. A uniformly channel doping concentration is assumed with a value of 1×10^{18} / cm^3 [25, 26].

In our proposed structure, the gate material is based on Graphene which is justified by several advantages amongst we cite the Graphene transparency reducing the shadowing effect taking place at the level of RADFET sensitivity in case of polysilicon gate material. In addition, the Graphene work function tunnability offers more flexibility for the efficient design of RADFET devices. Moreover, the Graphene may play a vital role in enhancing the transport mechanisms if adopted as a channel body due to its high conductivity [27].

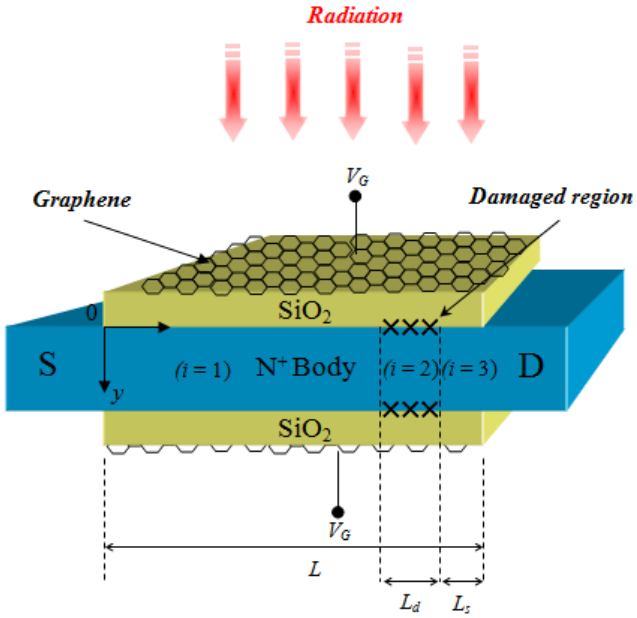


Fig.V.1. Schematic view of the proposed JL-DGG RADFET Dosimeter.

Holes produced in the SiO_2 layer travel and then are captured at the Si/SiO_2 boundary when a positive bias voltage is applied to the gate, generating an observable threshold-voltage change, as illustrated in **Figure V.1**. At the interface of the gate oxide and channel, the irradiation causes trap charges, which are known as fixed charges [28]. **Table V.1** summarizes the design parameters assigned to the schematic view.

Table V.1 Reference parameters of the proposed dosimeter.

Parameters	Value
Channel length, L_{ch} *(nm)*	40
Body thickness, t_b (nm)	20
Doping concentration, N_d (cm^{-3})	10^{18}
Gate oxide thickness, t_{ox} (nm)	2
Gate work function, ϕ_{Gr} (eV)	4.5

V.3. Analytical modeling methodology

Figure V.1 depicts the configuration of the JL DGG RADFET including a fixed charge zone at the boundary between the channel and gate oxide areas. In more detail, the channel is partitioned into three regions according to the localized charges' position [29]. The gate is made by monolayer graphene material.

A parabolic form is adopted to approximate the body potential $\phi(x, y)$ in the form $\phi(x, y) = \phi_s + a_1(x)y + a_2(x)y^2$, where ϕ_s denotes the surface potential. The Poisson equation that includes the localized charges effect is expressed as follows:

$$\frac{\partial^2 \phi_S}{\partial x^2} + \frac{\partial^2 a_1}{\partial x^2} y_d + \frac{\partial^2 a_2}{\partial x^2} y_d^2 + 2a_2 = -\frac{qN_{sub}}{\varepsilon_{Si}} \quad (1)$$

where y_d denotes the width of depletion zone and N_{sub} represents doping concentration in the body. By applying the appropriate boundary conditions taking into account the electrical flux continuity at the border and using Gauss' theorem to the middle of the channel body ($y_d/2$), a_1 and a_2 are defined as functions of y_d and ϕ_s [30]. The width of depletion y_d becomes $t_{si}/2$ when the threshold voltage is attained. Hence, the boundary conditions can be expressed by:

$$\begin{cases} \phi(x, y_d) = V_C \\ \partial\phi(x, y)/\partial y\big|_{y=0} = C_{ox}/\varepsilon_{Si}\,(\phi_S(x) - V_G + V_{FB}) \\ \partial\phi(x, y)/\partial y\big|_{y=t_{Si}/2} = 0 \end{cases} \quad (2)$$

with V_C represents the potential at the middle of the channel.

The corresponding second-order differential equation is developed by imposing both of these conditions to (1):

$$\frac{\partial^2 \phi_S}{\partial x^2} - k^2 \phi_S = -k^2 \phi_i \quad (3)$$

$$\phi_i = V_G + \frac{qN_{sub} t_{Si}}{2C_{ox}} - V_{FBi} \; (i = 1,2 \; and \; 3) \quad (4)$$

$$k^2 = \frac{8C_{ox}}{t_{Si}(4\varepsilon_{Si} + C_{ox} t_{Si})} \quad (5)$$

It is to note that the index values i = 1, 2, and 3 refer to the three distinct zones. V_{FBi} represents the flat band voltage and is expressed as follows:

- For i=2 we have $V_{FBi} = V_{FB0} - qN_f/C_{ox}$
- For i=1, 3 we have $V_{FBi} = V_{FB0}$

where V_{FB0} stands for the safe area's flat-band voltage and N_f represents the density of the localized charges. The general solution of (3) is expressed by

$$\phi_{S,i}(x) = b_i e^{kx} + c_i e^{-kx} + \phi_i \quad (6)$$

Fulfilling the continuity condition of the electric flux at the edges amid the fresh area and the damaged area permits to calculate both parameters b_i and c_i. (See the Appendix).

The threshold condition arises when y_d attains its maximum value, and ϕ_s achieves its lower limit since the considered n-type body holds high doping values [31]. Applying $\partial\phi_s / \partial x = 0$; the minimum surface potential $\phi_{s,min}$ is formulated as

$$\phi_{S,\min} = 2\sqrt{b_i\, c_i} + \phi_i \quad (7)$$

As stated previously, the threshold voltage condition is achieved when $y_{d,max}$ equals to $t_{si}/2$. Consequently, $\phi_{s,min}$ in this case can be calculated by adopting $y_d = t_{si}/2$ in Gauss's law. Therfore, $\phi_{s,min}$ can be expressed as follows:

$$\phi_{S,\min} = V_C - \frac{qN_{sub}\, t_{Si}^2}{8\,\varepsilon_{Si}} \quad (8)$$

The threshold voltage expression can be acquired by incorporating (4) with (7) and (8).

$$V_{Ti} = V_{T0} - 2\sqrt{b_i c_i} - \frac{qN_f}{C_{ox}}\delta(i-2) \quad (9)$$

$$V_{T0} = V_{FB0} + V_C - \frac{qN_{sub}\, t_{Si}}{2\,C_{ox}} - \frac{qN_{sub} t_{Si}^2}{8\,C_{ox}} \quad (10)$$

where δ(i−2) is equal to 1 for i = 2 and equal to 0 for i = 1, 3 [32]. The parameters b_i and c_i are quadratic equations of V_{Ti}. Therefore, by solving (9), the threshold voltage V_{Ti} is provided, which is another quadratic equation. By taking the highest value among different threshold voltages

associated to distinct areas, V_T is then chosen and determined appropriately. Formulas (11), (12), and (13) show each solved V_T equation, where the related constants are stated in the Appendix.

$$V_{T1} = \frac{V_{T0} + 2(\alpha\beta_1^* + \alpha^*\beta_1) - \sqrt{(V_{T0} + 2(\alpha\beta_1^* + \alpha^*\beta_1))^2 - (4\alpha\alpha^* - 1)(4\beta_1\beta_1^* - V_{T0}^2)}}{(1 - 4\alpha\alpha^*)} \quad (11)$$

$$V_{T2} = \frac{V_{T0} - \frac{qN_f}{C_{ox}} + 2(\alpha\beta_2^* + \alpha^*\beta_2) - \sqrt{(V_{T0} + 2(\alpha\beta_2^* + \alpha^*\beta_2))^2 - (4\alpha\alpha^* - 1)\left[4\beta_2\beta_2^* - (V_{T0} - \frac{qN_f}{C_{ox}})^2\right]}}{(1 - 4\alpha\alpha^*)} \quad (12)$$

$$V_{T3} = \frac{V_{T0} + 2(\alpha\beta_3^* + \alpha^*\beta_3) - \sqrt{(V_{T0} + 2(\alpha\beta_3^* + \alpha^*\beta_3))^2 - (4\alpha\alpha^* - 1)(4\beta_3\beta_3^* - V_{T0}^2)}}{(1 - 4\alpha\alpha^*)} \quad (13)$$

V.4. Background of genetic algorithms

Originally, genetic algorithms (GAs) have been recognized as an optimization approach, inspired from natural evolution thanks to the employment of several genetic operators. In this framework, population evolves iteratively with the aim of reaching satisfactory stable solutions with respect to a given figure of merit [33]. A general framework describing the execution of different stages is depicted in **Figure V.2**.

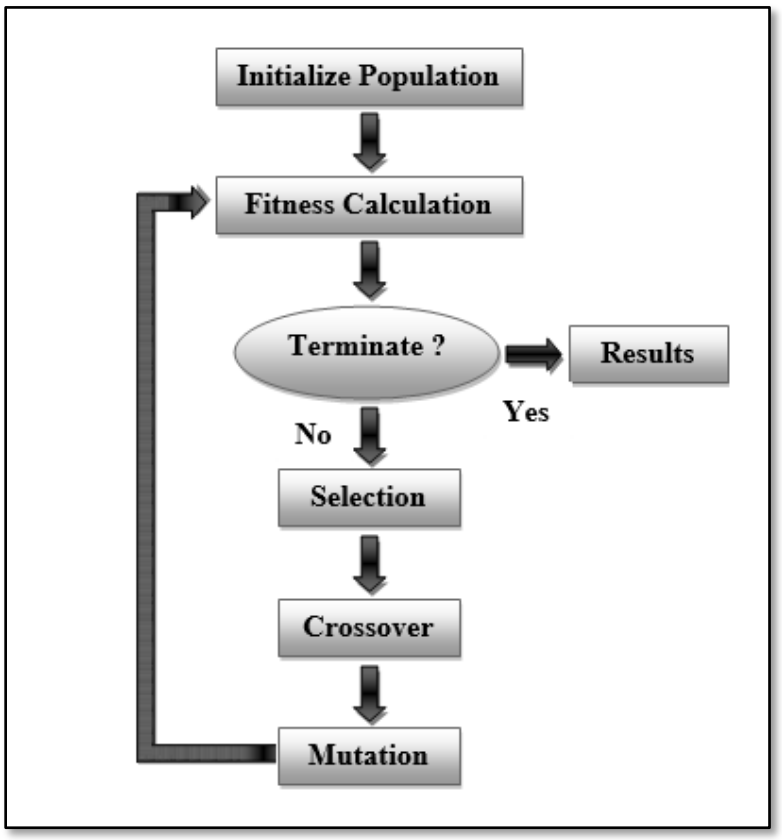


Fig.V.2. Main steps performed by GA.

Multi-objective genetic algorithms can be defined as a dynamic strategy to seek the optimum solution of restricted and unrestricted issues, where various figures of merit must be analyzed jointly [22, 27, 34–36]. These approaches ensure the availability of an acceptable range of solutions according to the area of operation and thus offer a quick implementation and a large amount of knowledge. The basis of genetic algorithms derives from the natural progression of species. In other words, they randomly pick individuals and develop new generations named offspring, starting from an initial population. For various generations, this process is repeated to fulfill the stopping conditions that provide at least a near optimal solution. **Figure V.3** provides a basic flowchart for MGAs.

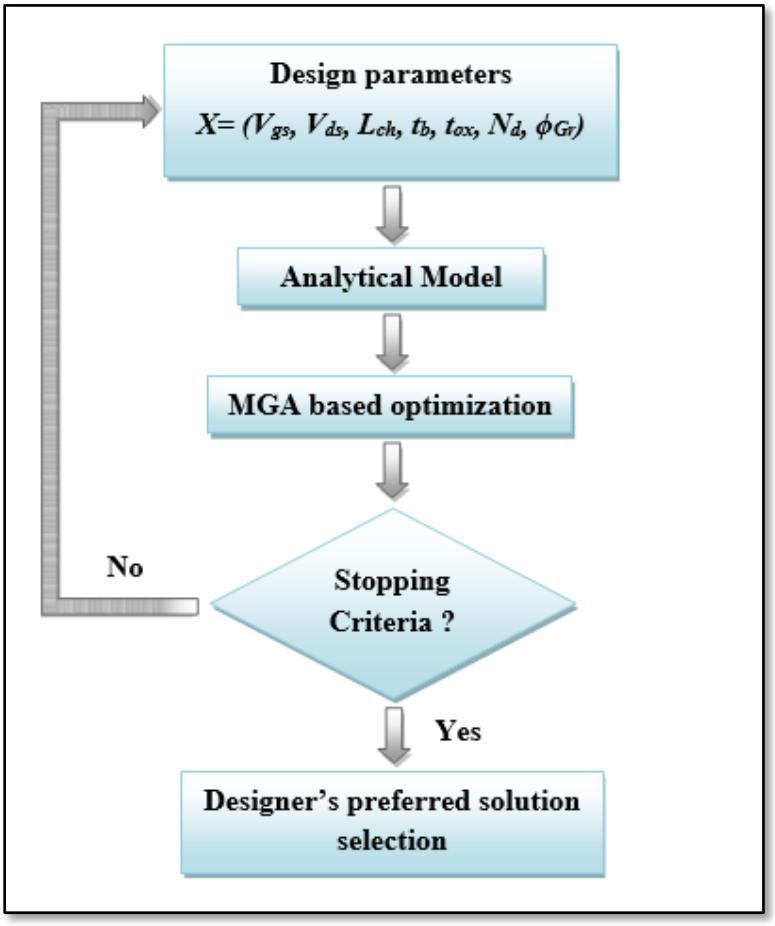


Fig.V.3. A representative flowchart of MGA.

The essential operations implemented on such paradigm to enable the process progression are selection, crossover and mutation. The selection concerns the transfer of a portion of the best individuals to the next generation without going through additional modifications. The crossover is preserved for generating new offspring individuals based on parent solutions, which is crucial for gaining further diversity in the population. The mutation can be seen as a kind of local modification within the solution structure in order to intensify search at the vicinity of the provided solution. The solution encoding used within the MGA terminology, refers to the geometrical and electrical parameters of the structure, like channel length, oxide thickness, graphene work function and doping density, which have a vital role in deciding the significant outcomes of the DGG JL RADFET [37]. By adopting a genetic based technique, it would be possible to benefit from the many associated advantages including: flexibility of encoding and implementation, reasonable computation time and feasibility of adaptation to parallel platforms. All these aspects are mandatory required for the design of deeply scaled electronic devices.

V.5. Simulation experiments and discussions

Figure V.4 represents the surface potential distribution across the channel length for various values of localized charge density and different L_d lengths based on the suggested model.
It can be seen that negative (positive) charges at Si−SiO_2 boundary shift downward (upward) the surface potential, where higher discrepancies between potential barrier curves can be detected with the expansion of the damaged zone. This means that at negative charges, as seen in **Figure V.5**, a larger V_{th} change will be noticed.

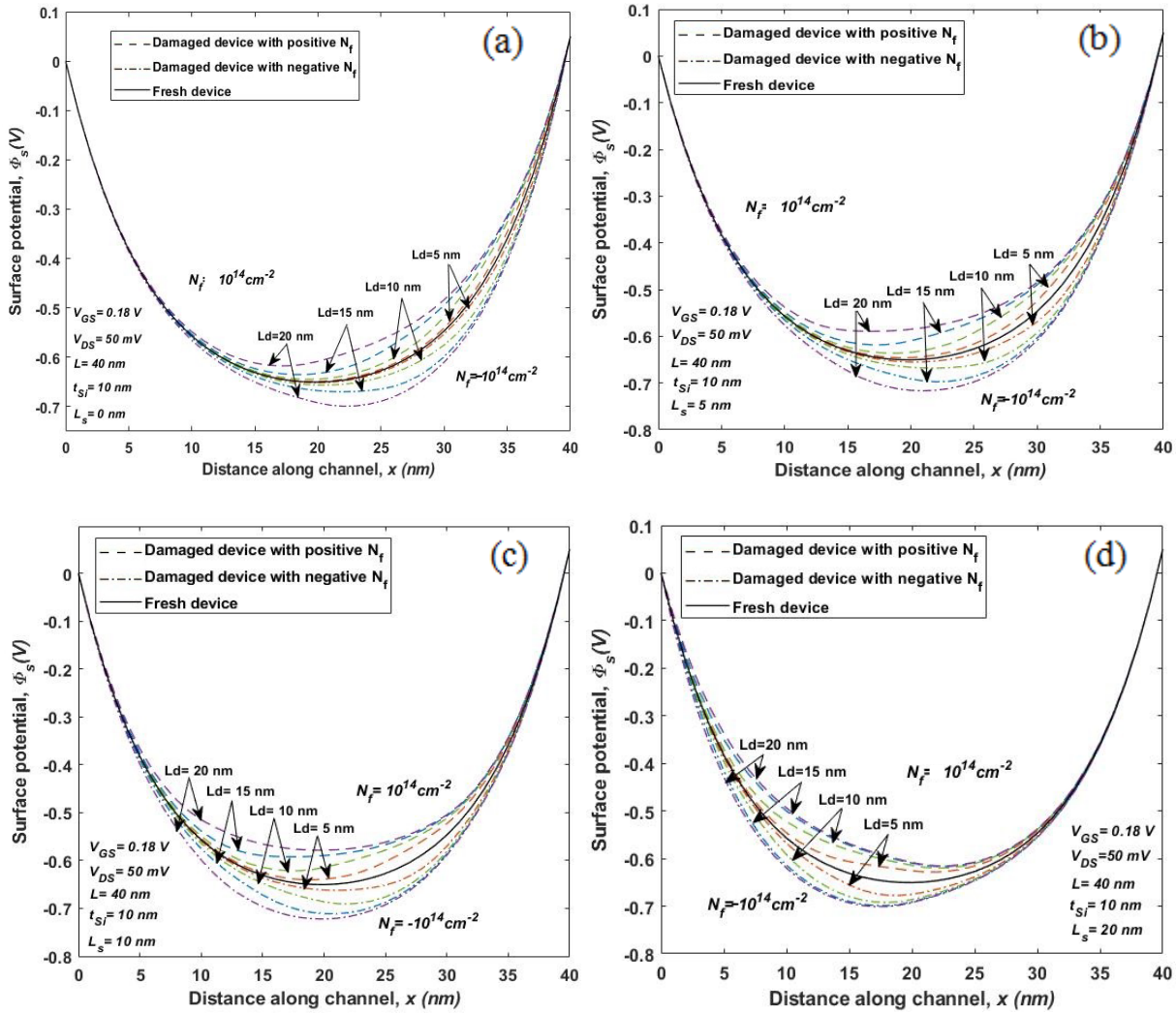


Fig.V.4. Variation of surface potential along the channel with different values of damaged zone length for a) L_S= 5 nm, b) L_S =10 nm, c) L_S =15 nm and d) L_S=20 nm, respectively, with positive and negative N_f.

Because more deformation is detected in the surface potential, as the affected zone becomes largēr, further shift in V_{th} is also observed. Besides, the more the undamaged zone increases the surface potential shifts to the source side.

The sensitivity is expressed as follows:

$$S = \frac{\Delta V_{th}}{D} \qquad (14)$$

where S and ΔV_{th} denote the RADFET sensitivity and the alteration in threshold voltage. The absorbed dose is symbolized by D. From equation (14), it is clear that a higher threshold voltage alteration yields a higher sensitivity.

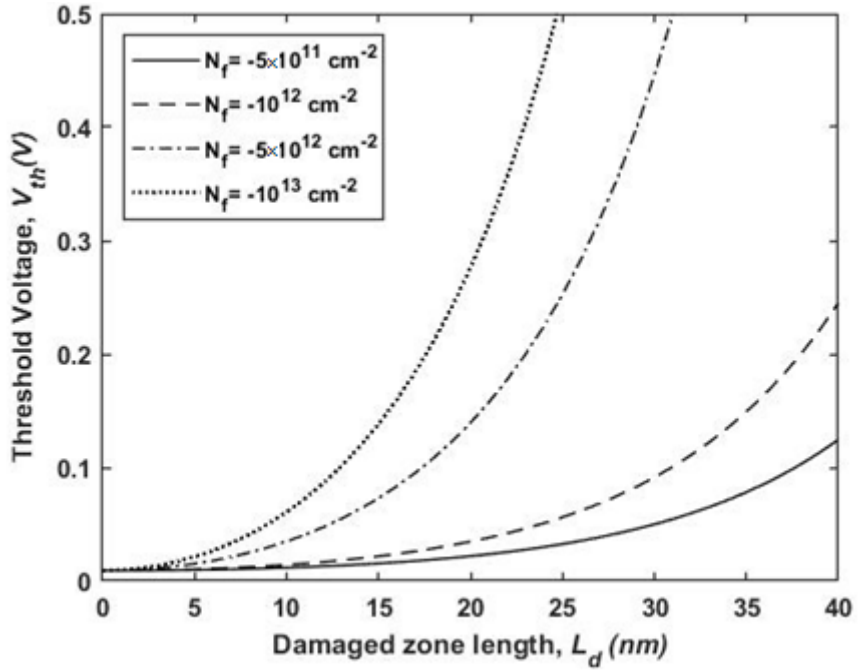


Fig.V.5. Threshold voltage variation versus damaged zone length for different localized charges densities (N_f= -5×10^{11}cm^{-2}, N_f = -10^{12}cm^{-2}, N_f = -5×10^{12}cm^{-2} and N_f =-10^{13}cm^{-2}).

As provided in **Figure V.6**, the device shows a significant I_{on}/ I_{off} ratio. It highlights that the electrostatics reliability of the design is satisfactory.

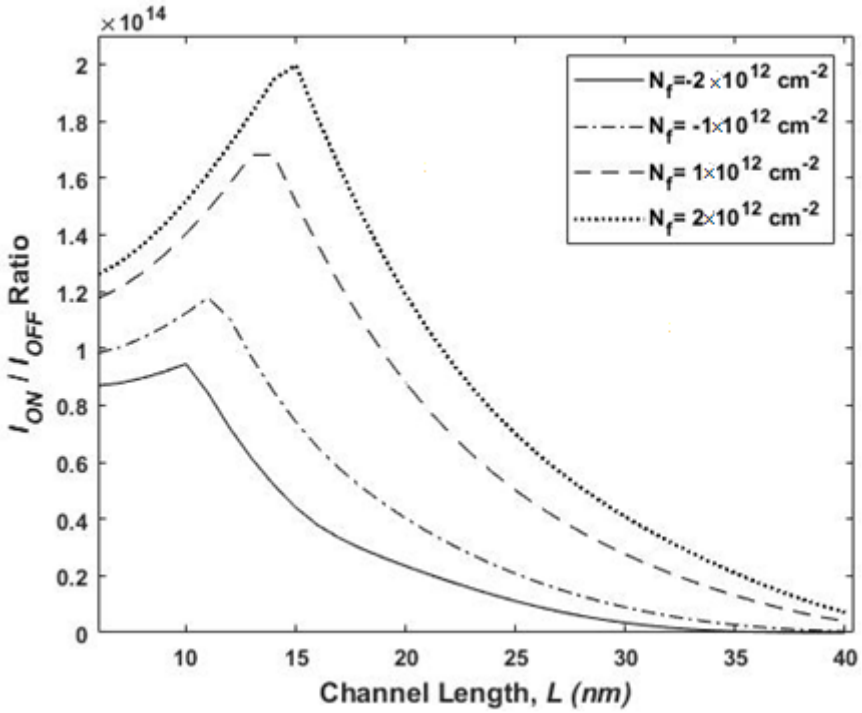


Fig.V.6. I_{ON}/I_{OFF} ratio variation versus the channel length for different traps densities (N_f= -2×10^{12} cm^{-2}, N_f = -1×10^{12}cm^{-2}, N_f = 1×10^{12}cm^{-2} and N_f =2×10^{12}cm^{-2}).

Figure V.7 illustrates the surface potential variation across the channel with diverse values of damaged zone length and different body silicon thicknesses.

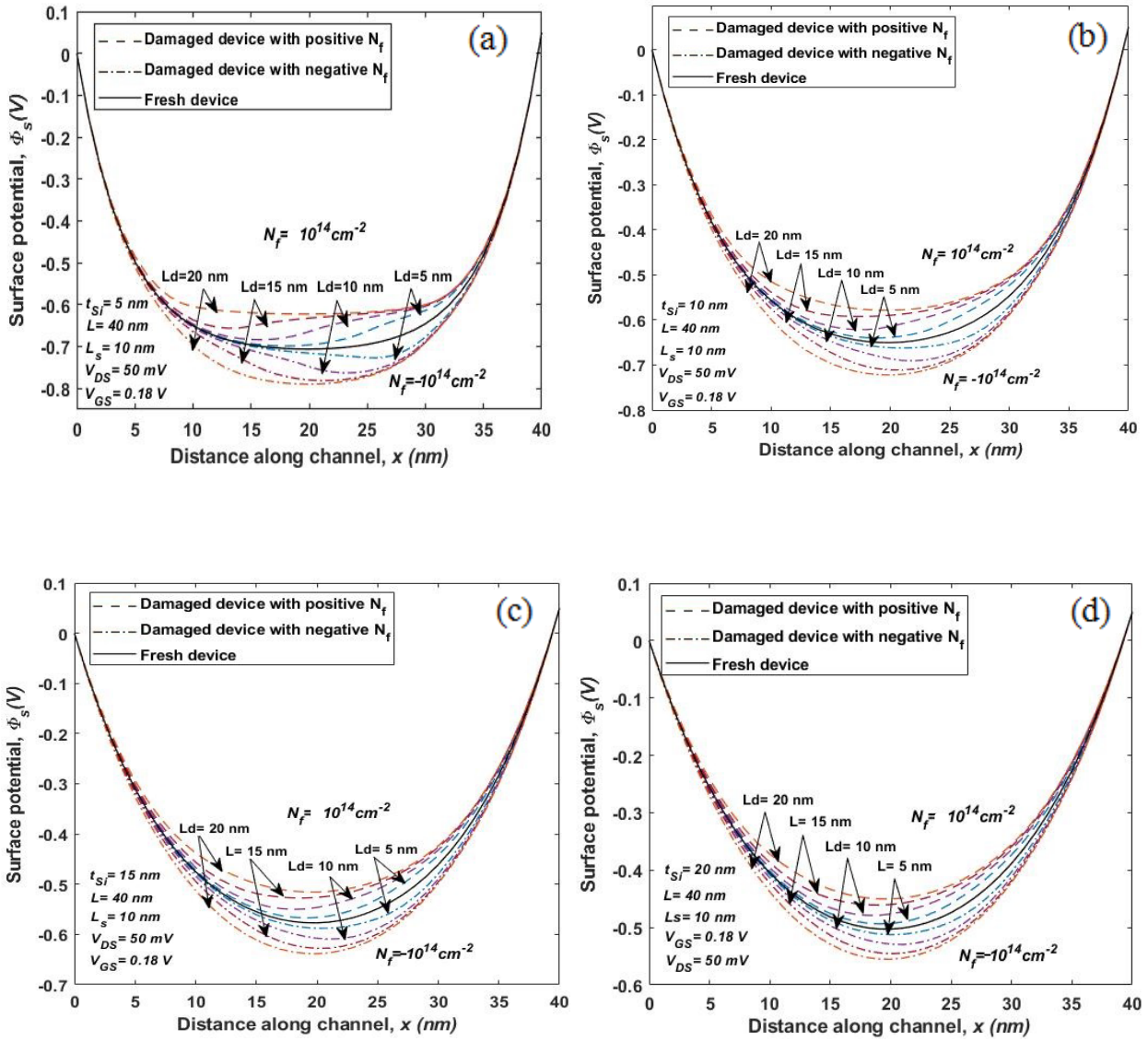


Fig.V.7. Variation of surface potential along the channel with different values of channel thickness for a) t_{Si}= 5 nm, b) t_{Si} =10 nm, c) t_{Si} =15 nm and d) t_{Si}=20 nm, for N_f=10^{14}cm^{-2}, respectively.

In **Figure V.7**, it is confirmed that the surface potential is more susceptible to localized charges when t_{Si} is 5 nm than when t_{Si} is 20 nm. We can disclose that the positive localized charges aggravate the short-channel impact. In comparison, as the body is thick, these impacts are more extreme, since the quotient channel length to body thickness is lower, while the channel is shortened by similar length. From the other side, the channel inversion is avoided by negative localized charges and thus, V_{th} is raised; as the body is smaller, this effect is significant.

It is noticed in **Figure V.8** that V_{th} is also influenced by various body thicknesses with similar charges. The threshold voltage is diminished as t_{Si} grows, and the form of the variance of V_{th} is adjusted accordingly. This fact is due to the surface potential variation with respect to the localized charges when the Silicon body becomes thicker. In thinner devices, the influence of localized charges on V_{th} is thus important. Nevertheless, for similar thicknesses, the tendency of V_{th} degradation is identical. This means that the SCEs mainly rely on the amount of trapped charges for various body thicknesses.

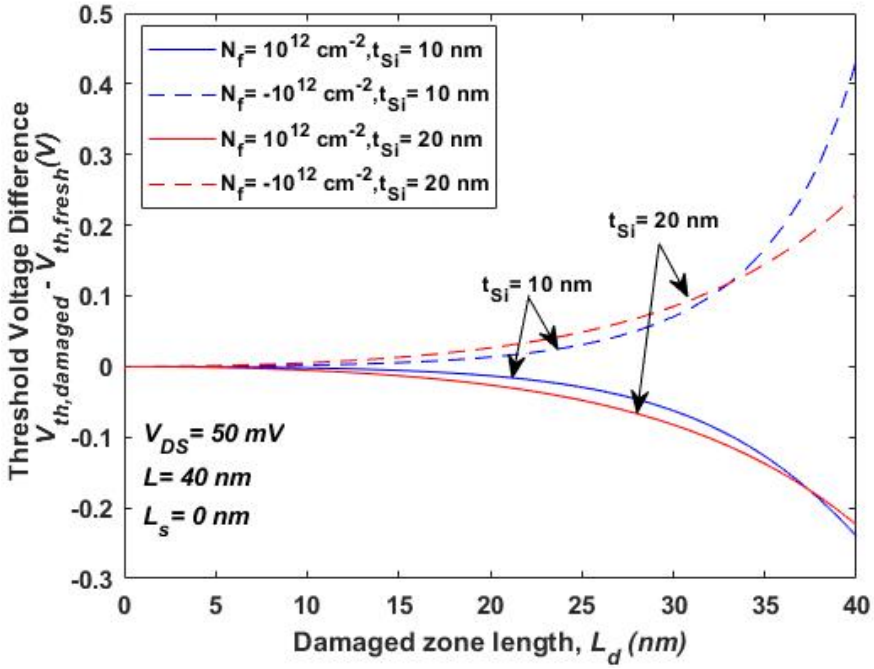


Fig.V.8. Threshold voltage difference variations versus damaged zone length with positive/negative localized charges.

Figure V.9 is introduced to explain the dependence of threshold voltage on oxide thickness. With increasing oxide thickness, it is observed that the threshold voltage increases, which means a rise in the device sensitivity. We can explain such behavior by the inverse proportionality relating the threshold voltage and oxide capacitance, which augments with oxide thickness reduction.

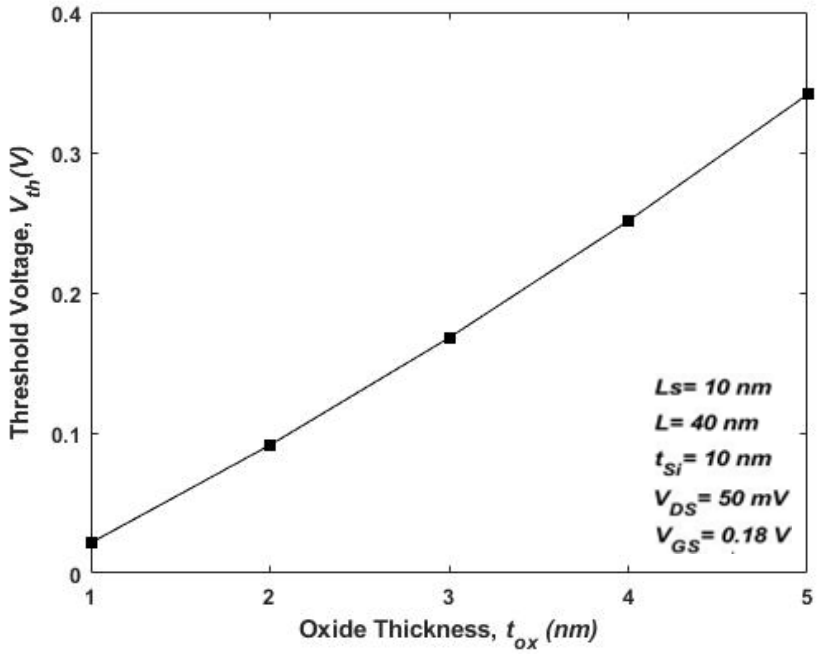


Fig.V.9 Threshold voltage variation as a function of oxide thickness.

Latest experiments have demonstrated that by imposing an electric field or chemical doping, the work function of graphene can be regulated between bounds 3.5 and 5.16 eV [38, 39]. Furthermore, the graphene work function is often influenced by the number of graphene layers [40]. In this work, we just consider the influence on the performance of DG JL RADFET of the adjusting monolayer graphene work function. As seen in **Figure V.10**, with the increase of the graphene work function, the surface potential decreases, which is primarily due to the raise of the Schottky barrier (φ_B).

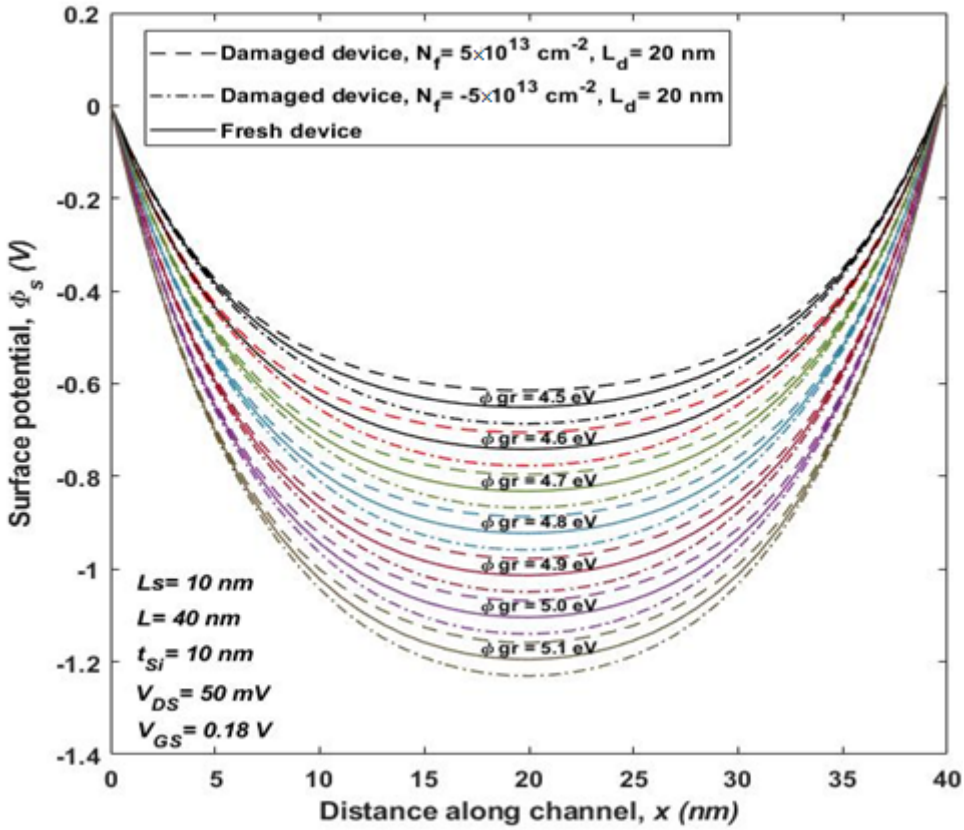


Fig.V.10 Surface potential variation along the channel for different values of graphene work function.

Because more deformations in ϕ_s are revealed in **Figure V.10**, this means that larger V_{th} changes will be noted in **Figure V.11**. Besides, the more the damaged zone increases the more the threshold voltage increases.

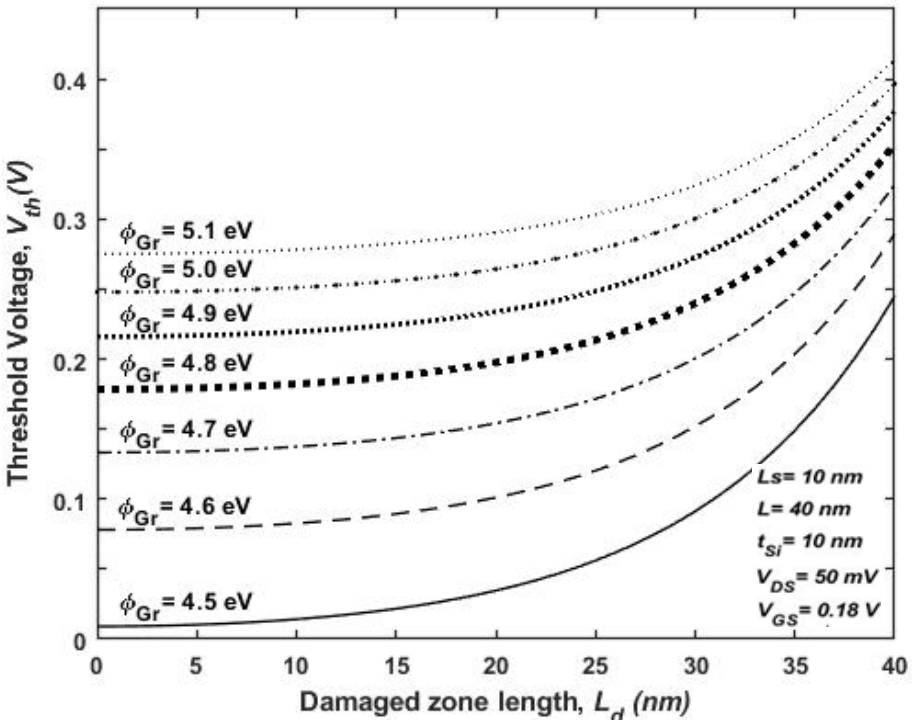


Fig.V.11 Threshold voltage variations versus damaged zone length for different values of graphene work function.

V.5.1. Optimized JL-DGG dosimeter using MGA' approach

Because of low implementation costs and high flexibility delivered by MGA-based method for multi-objective design, MGA can be employed to exploit and boost the DG JL G-RADFET sensitivity performances. A multi-objective optimization problem is in general specified by a solution search space, several objective functions and an ensemble of constraints. In our scenario, we adopt three objective functions that define the JL-DGG RADFET in terms of output sensitivity and reliability.

$$F(X) = \begin{pmatrix} F_1(X) \\ F_2(X) \\ F_3(X) \end{pmatrix} \qquad (19)$$

where F_1(X)= V_{th}, F_2(X)= S and F_3(X)= I_{ON}/I_{OFF}, X denotes the variables vector X= (V_{gs},V_{ds}, L_{ch}, t_b, t_{ox}, N_d, ϕ_{Gr}) including 7 components.

With regard to the following targets, the JL-DGG -RADFET is optimized following the provided rules:

- Maximize the first objective function, $F_1(X)$.
- Maximize the second objective function, $F_2(X)$.
- Maximize the third objective function, $F_3(X)$.

The ensemble of constraints makes it possible to specify the requirements on the search space to be fulfilled by the design parameters.

For the considered scenario, these constraints are provided using the following rules:

- $g_1(x): x \in [x_{i\min}, x_{i\max}],\ x_i \in X$ (Inside a given set, each design variable should be limited.).

- $g_2(x): L_1 + L_2 + L_s \leq L.$

- $g_3(x): L_1 + L_2 + L_s = 40\ nm.$

Every vector is binary-coded in a more adequate formulation and is named chromosome [22]. With a view to imitating nature, the steps of crossover and mutation are planned to accelerate the convergence of the populations towards near optimal solutions. Comparing recent chromosomes in the next stage of selection and favoring the removal of those with low fitness values allow boosting the solutions quality. The configuration parameters adopted during MGA simulation are recapitulated in **Table V.2**, where the used stopping criteria are either by reaching the total number of generations (1000 iterations) or by satisfying a given tolerance threshold (10^{-6}).

Table V.2 Configuration parameters adopted during MGA simulation [22, 27, 34–36]

Parameters	Value
Number of design variables	7
Population size	1000
Maximum number of generations	100
Selection type	Tournament
Crossover type	Scattered
Mutation type	Adaptative feasible migration
Crossover rate	0.8
Migration rate	0.2
Pareto front population fraction	0.5

The solution of our MGA-based optimization, unlike mono-objective optimization, is not determined by a unique solution, but instead by a set of compromise solutions (non-dominated solutions), recognized as Pareto front solutions where each point is connected to a well-defined vector X combination. We have selected three specified points from the non-dominated solutions to test our optimization, and the related objective functions of DGG JL RADFET sensitivity efficiency are presented in **Table V.3**.

Table V.3 Optimized DGG JL RADFET design parameters.

Design parameters	***Case 1***	***Case 2***	***Case 3***
V_{GS} *(V)*	0.22	0.2	0.18
V_{DS} *(V)*	0.05	0.08	0.05
Channel length, L_{ch} (nm)	40	40	40
Gate oxide thickness, t_{ox} (nm)	5	5	2
Body thickness, t_b (nm)	20	20	20
Body doping concentration, N_d (cm^{-3})	5×10^{17}	10^{17}	10^{18}
Gate work function	5	5.1	4.8
Objective functions			
V_{th} (V)	1	**1.01**	0.77
S (mV/Gy)	**2.11**	2.11	1.65
I_{ON}/I_{OFF}	2.83×10^{13}	3.45×10^{13}	$\mathbf{3.53\times10^{13}}$

The maximal and minimal objective functions ($F_1(x)$ and $F_2(x)$) are correlated with two extreme situations (case 1 and case 3) in the problem space respectively, and the third point essentially corresponds to the mono-objective optimizing. Each solution of Pareto front is therefore non dominated which means that at minimum one objective cannot be decreased without altering other objectives. In addition, the findings acquired from Pareto front allow the designer to provide a detailed overview of the compromises in the design with regard to various figures of merit.

We consider now mono-objective-based optimization in order to achieve a balance among both sensitivity and reliability. In this case, the ultimate mono-objective model can be defined by provided weighting factors based on the weighted sum method.

$$F(X) = w_1 V_{th} + w_2 \frac{\mathrm{I_{ON}}}{\mathrm{I_{OFF}}} + w_3 \mathrm{S} \qquad (21)$$

where w_i *(i = 1-3)* can be taken equal to 1/3.

Table V.4 underlines some of the calculated results. It should be noted that the efficiency of the optimized JL DGG RADFET is superior with regard to I_{ON}/I_{OFF} ratio and threshold voltage than those of state-of-the-art RADFETs. If we assume a dose value equal to 104 Gy in order to evaluate various criteria of the suggested RADFET and the standard DG RADFETs, **Table V.4** is obtained, where it shows that the performance measures of the JL DGG RADFET outperform those of standard RADFETs.

Table V.4 Benchmarking table.

Design parameters	**DG RADFET [5]**	**JL DG RADFET [5]**	**Proposed JL-DGG RADFET**	**Optimized JL-DGG RADFET**
V_{GS} *(V)*	0.18	0.18	0.18	0.2
V_{DS} *(V)*	0.05	0.05	0.05	0.08
Channel length, L_{ch} (nm)	40	40	40	40
Gate oxide thickness, t_{ox} (nm)	2	2	2	5
Body thickness, t_b (nm)	20	20	20	20
Body doping concentration, N_d (cm^{-3})	10^{18}	10^{18}	10^{18}	10^{17}
Gate work function	4.5 (polysilicon)	4.5 (polysilicon)	4.5 (Graphene)	5.1 (Graphene)
Objective functions				
V_{th} (V)	0.83	0.85	0.92	1.01
S (mV/Gy)	1.68	1.92	1.97	2.11
I_{ON}/I_{OFF}	1.5×10^{9}	3.5×10^{12}	4×10^{12}	3.45×10^{13}

From this table, it is clearly shown that, compared to the traditional JL-DGG RADFET structure, our optimized configuration provides better sensitivity and reliability quality, which renders it a possible candidate to resolve some sever challenges.

V.6. Conclusion

The Junction-less Double Graphene Gate RADFET analytical modeling was investigated as an efficient framework to attend high performance for radiation sensing applications. Changes in threshold voltage, surface potential and I_{ON}/I_{OFF} ratio were studied for various values of localized charges. In comparison to the standard Double Gate RADFET, threshold voltage, sensitivity and I_{ON}/I_{OFF} ratio of the suggested RADFET were evaluated, where an improvement in terms of threshold voltage, I_{ON}/I_{OFF} ratio and sensitivity has been recorded. In addition, the derived analytical models have served as fitness functions for the MGA technique. The optimized configuration exhibits higher responses with respect to some state of the art DG RADFETs. As future perspectives, additional efforts may be focused on the consideration of other geometrical structures, in addition to the investigation of innovative channel materials such as InGZnO and SiC.

APPENDIX

By applying the potential and the electrical flux continuity at the edges, the solution for (6) can be achieved. Using $\phi_{s,i}(x) = \phi_{s,i+1}(x)$, $\phi_{s,i}(x) = \phi_{s,i+1}(x)$, $\phi s(0) = 0$ and $\phi_s(L) = V_D$ at all borders, six equations can be deduced [31]. By manipulating the given equations (A-1–A- 15), both parameters b_i and c_i can be derived using (A-1)–(A-6). We consider in the aforementioned equations, $L_1 = L-L_d-L_s$ and $L_2 = L-L_s$. It was already shown that $\phi_1-\phi_2$ is $-qN_f/C_{ox}$, and ϕ_1 is obtained by adding V_T and constants. Consequently, b_i and c_i can be written as $\alpha V_T + \beta$ with α and β are constants as indicated in (A-7)–(A-15), and the asterisk (*) constants can be in turn deduced by substituting k with $-k$. By replacing (A-7) into (9), the resolution of three quadratic expressions leads to V_T as given in (11), (12), and (13).

$$b_1 = \frac{V_D + \phi_1(e^{-kL} - 1) + (\phi_1 - \phi_2)\left[\cosh k(L - L_2) - \cosh k(L - L_1)\right]}{2\sinh kL} \quad \text{(A-1)}$$

$$b_2 = \frac{V_D + \phi_1(e^{-kL} - 1) + (\phi_1 - \phi_2)\left[\cosh k(L - L_2) - e^{-kL}\cosh kL_1\right]}{2\sinh kL} \quad \text{(A2)}$$

$$b_3 = \frac{V_D + \phi_1(e^{-kL} - 1) + (\phi_1 - \phi_2)e^{-kL}\left[\cosh k\,L_2 - \cosh kL_1\right]}{2\sinh kL} \quad \text{(A3)}$$

$$c_1 = \frac{-V_D - \phi_1(e^{kL} - 1) - (\phi_1 - \phi_2)\left[\cosh k(L - L_2) - \cosh k(L - L_1)\right]}{2\sinh kL} \quad \text{(A-4)}$$

$$c_2 = \frac{-V_D - \phi_1(e^{-kL} - 1) - (\phi_1 - \phi_2)\left[\cosh k(L - L_2) - e^{kL}\cosh kL_1\right]}{2\sinh kL} \quad \text{(A5)}$$

$$c_3 = \frac{-V_D - \phi_1(e^{kL} - 1) - (\phi_1 - \phi_2)e^{kL}\left[\cosh k\,L_2 - \cosh kL_1\right]}{2\sinh kL} \quad \text{(A6)}$$

$$\begin{cases} b_1 = \alpha V_T + \beta_1 \\ b_2 = \alpha V_T + \beta_2 \\ b_3 = \alpha V_T + \beta_3 \\ c_1 = \alpha^* V_T + \beta_1^* \\ c_2 = \alpha^* V_T + \beta_2^* \\ c_3 = \alpha^* V_T + \beta_3^* \end{cases} \quad \text{(A-7)}$$

$$\alpha = \frac{e^{-kL} - 1}{2\sinh kL} \quad \text{(A-8)}$$

$$\alpha^* = \frac{1 - e^{kL}}{2\sinh kL} \quad \text{(A-9)}$$

$$\beta_1 = \frac{V_D + (\frac{qN_{sub}t_{Si}}{2C_{ox}} - V_{FB0})(e^{-kL} - 1) - \frac{qN_f}{C_{ox}}\left[\cosh k(L - L_2) - \cosh k(L - L_1)\right]}{2\sinh kL} \quad \text{(A-10)}$$

$$\beta_2 = \frac{V_D + (\frac{qN_{sub}t_{Si}}{2C_{ox}} - V_{FB0})(e^{-kL} - 1) - \frac{qN_f}{C_{ox}}\left[\cosh k(L - L_2) - e^{-kL}\cosh kL_1\right]}{2\sinh kL} \quad \text{(A-11)}$$

$$\beta_3 = \frac{V_D + (\frac{qN_{sub}t_{Si}}{2C_{ox}} - V_{FB0})(e^{-kL} - 1) - \frac{qN_f}{C_{ox}}e^{kL}\left[\cosh kL_2 - \cosh kL_1\right]}{2\sinh kL} \quad \text{(A-12)}$$

$$\beta_1^* = -\frac{V_D + (\frac{qN_{sub}t_{Si}}{2C_{ox}} - V_{FB0})(e^{kL} - 1) - \frac{qN_f}{C_{ox}}\left[\cosh k(L - L_2) - \cosh k(L - L_1)\right]}{2\sinh kL} \quad \text{(A-13)}$$

$$\beta_2^* = -\frac{V_D + (\frac{qN_{sub}t_{Si}}{2C_{ox}} - V_{FB0})(e^{kL} - 1) - \frac{qN_f}{C_{ox}}\left[\cosh k(L - L_2) - e^{kL}\cosh kL_1\right]}{2\sinh kL} \quad \text{(A-14)}$$

$$\beta_3^* = -\frac{V_D + (\frac{qN_{sub}t_{Si}}{2C_{ox}} - V_{FB0})(e^{kL} - 1) - \frac{qN_f}{C_{ox}} e^{kL}\left[\cosh kL_2 - \cosh kL_1\right]}{2\sinh kL} \qquad \text{(A-15)}$$

References

[1] A. Kumar, J. N. Roy, IEEE Transactions on Electron Devices, 66(8), 3640-3645, (2019).

[2] A. Dubey, M. Gupta, R. Narang, M. Saxena, In 2016 IEEE International Nanoelectronics Conference (INEC) (pp. 1-2). IEEE, (2016, May).

[3] D. Arar, F. Djeffal, T. Bentrcia, M. Chahdi, physica status solidi (c), 11(1), 65-68, (2014).

[4] A. Dubey, M. Gupta, R. Narang, M. Saxena, In 2018 4th International Conference on Devices, Circuits and Systems (ICDCS) (pp. 117-120). IEEE, (2018, March).

[5] A. Dubey, A. Singh, R. Narang, M. Saxena, M. Gupta, Modeling and simulation of junctionless double gate radiation sensitive FET (RADFET) dosimeter. IEEE Transactions on Nanotechnology, 17(1), 49-55, (2017).

[6] F. Meddour, A. Meddour, M. A. Abdi, M. Amir, 2018 International Conference on Communications and Electrical Engineering (ICCEE) IEEE, 1-4, (2018).

[7] G. F. Knoll, New York, NY, USA: Wiley, (1989).

[8] A. Holmes-Siedle and L. Adams, Radiation Phys. Chem., 28, 235–244, (1986).

[9] J. M. Benedetto and H. E. Boesch, Sci., 31 (6), 1461– 1466, (1984).

[10] A. Holmes-Siedle, (121), 169–179, (1974).

[11] G. Risti´c, S. Golubovi´c, and M. Pejovi´c, Sens. Actuators A, Phys, 51, 153–158, (1996).

[12] A. Kelleher, N. McDonnell, B. O'Neill, L. Adams, and W. Lane, Sens. Actuators A, vol. 37/38, 370–374, (1993).

[13] D. J. Gladstone, X. Q. Lu, J. L. Humm, H. F. Bowman, and L. M. Chin, Med. Phys, 21, 1721–1728, (1994).

[14] F. Pezzimenti, H. Bencherif, A. Yousfi, L. Dehimi, Solid-State Electronics, 161, 107642, (2019).

[15] A. Enright, Master Thesis, National University of Ireland, Cork, Ireland, (1996).

[16] B. O'Connell, C. Conneely, C. McCarthy, J. Doyle, W. Lane, and L. Adams, IEEE Trans. Nucl. Sci., 45 (6), 2689–2694, (1998).

[17] G. Risti´c, A. Jakˇsi´c, and M. Pejovi´c, Sens. Actuators A, Phys., 63, 129–134, (1997).

[18] D. J. Gladstone, X. Q. Lu, J. L. Humm, H. F. Bowman, and L. M. Chin, Med. Phys., 21, 1721–1728, (1994).

[19] Y. Shi, K.K. Kim, A. Reina, M. Hofmann, L.-J. Li, J. Kong, ACS Nano 4 2689–2694, (2010).

[20] Y.-J. Yu, Y. Zhao, S. Ryu, L.E. Brus, K.S. Kim, P. Kim, Nano Lett. 9, 3430–3434, (2009).

[21] H. Hibino, H. Kageshima, M. Kotsugi, F. Maeda, F.Z. Guo, Y. Watanabe, Phys. Rev. B. 79, 125437, (2009).

[22] H. Bencherif, L. Dehimi, F. Pezzimenti, A. Yousfi, M. A. Abdi, L. Saidi, F. G. Della Corte, Optik, 223, 165346, (2020).

[23] H. Bencherif, L. Dehimi, F. Pezzimenti, F. G. Della Corte, Optik, *182*, 682-693, (2019).

[24] T. Joyce, J. M. Herrmann, 744, 27–51, (2018).

[25] J. H. Woo, J. M. Choi, Y. K. Choi, IEEE transactions on electron devices, 60 (9), 2951-2955, (2013).

[26] Kang, H., Han, J. W., & Choi, Y. K. IEEE electron device letters, 29(8), 927-930, (2008).

[27] M. A. Abdi, H. Bencherif, T. Bendib, F. Meddour, M. Chahdi, Sensors and Actuators A: Physical, 317, 112446, (2020).

[28] H. Bencherif, L. Dehimi, F. Pezzimenti, F. G. Della Corte, Applied Physics A, 125(5), 294, (2019).

[29] Y.-S. Jean and C.-Y. Wu, IEEE Trans. Electron Devices, 44, (3), 441–447, (1997).

[30] K. K. Young, IEEE Trans. Electron Devices, 36, 2, 399–402, (1989)

[31] H. Kang, J.-W. Han, and Y.-K. Choi, IEEE Electron Device Lett, 298, 927–929, 2008.

[32] J. H. Woo, J. M. Choi, Y. K Choi, IEEE transactions on electron devices, 60(9), 2951-2955, (2013).

[33] T. Bäck, Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms, Oxford University Press, United State, (1996).

[34] H. Bencherif, L. Dehimi, G. Messina, P. Vincent, F. Pezzimenti, F. G. Della Corte, Sens. Actuators A: Phys., 307, 112007.

[35] T. Bendib, F. Djeffal, IEEE Trans.Electron. Dev 58 3743–3750, (2011).

[36] H. Bencherif, L. Dehimi, F. Pezzimenti, G. De Martino, F.G. Della Corte, J. Electron. Mater, 48 (6), 3871–3880, (2019).

[37] D. S. Weile, E. Michielssen, IEEE Transactions on Antennas and Propagation, 45(3), 343-353, (1997).

[38] Y. Shi, K.K. Kim, A. Reina, M. Hofmann, L.-J. Li, J. Kong, ACS Nano, 4, 2689–2694, (2010).

[39] Y.-J.Yu,Y. Zhao,S. Ryu, L.E. Brus, K.S. Kim, P. Kim, NanoLett, 9, 3430–3434, (2009).

[40] H. Hibino, H. Kageshima, M. Kotsugi, F. Maeda, F.Z. Guo, Y. Watanabe, Phys.Rev.B. 79, 125437, (2009).

General conclusion

The focus of this PhD work was to investigate two different problems in radiation therapy, treatment planning, one pertaining to IMRT fluence map optimization and the other to RADFET Dosimeter enhancement. The following two sub-sections will summarize the pertinent conclusions for each of these optimization problems.

First, for the IMRT fluence map optimization, a mathematical modeling was presented, in which the clinical requirements for a treatment plan were transformed into a multiobjective optimization problem with multiple constraints. Then, the MGA was introduced to optimize the model. Lastly, a liver clinical example was tested. The results showed by pareto front confirm that an obtained set of non-dominated solutions were distributed uniformly. Then, the corresponding dose distribution of one of the solutions in the non-dominated solution set not only approached the expected dose distribution, but also satisfied the dose-volume constraints. It was indicated that the clinical requirements were better satisfied and that the planner could select the optimal treatment plan from the non-dominated solution set. With the method we proposed, the planner has no need for a trial and error process to find the optimum plan, so efficiency will be highly improved. As future perspectives, additional efforts may be focused on the consideration of other IMRT optimization problems, such beam angles selection and multileaf collimator MLC dose aperture optimization, in addition to the investigation of involving other optimization techniques.

Second, for the enhancement of radiation therapy quality QA and ensure a safe patient dose verification, the proposed Junction-less Double Graphene Gate RADFET analytical modeling was investigated as an efficient framework to attend high performance for radiation sensing applications. Changes in threshold voltage, surface potential and I_{ON}/I_{OFF} ratio were studied for various values of localized charges. In comparison to the standard Double Gate RADFET, threshold voltage, sensitivity and I_{ON}/I_{OFF} ratio of the suggested RADFET were evaluated, where an improvement in terms of threshold voltage, I_{ON}/I_{OFF} ratio and sensitivity has been recorded. In addition, the derived analytical models have served as fitness functions for the MGA technique. The optimized configuration exhibits higher responses with respect to some state of the art DG RADFETs. As future perspectives, additional efforts may be focused on the consideration of other geometrical structures, in addition to the investigation of innovative channel materials such as InGZnO and SiC.