國立清華大學 電機工程學系

實作專題研究成果摘要

Machine Learning for Device and Circuit Designs

機器學習應用於元件與電路設計

專題領域: 電子領域

組　　別: B537

指導教授: 黃敬源 (King-Yuen, Wong)

專題生: 陳聖凱、梁東麟

研究期間: 114 年 02 月 17 日 至 114 年 11 月 24 日，共 9 個月

摘要

在功率電子元件中，氮化鎵高電子遷移率電晶體(GaN HEMT)設計需要同時兼顧低導通電阻(Ron)、高擊穿電壓(BV)與低輸出電容(Coss)，這樣的多目標最佳化問題尤為複雜，設計上往往仰賴工程師反覆調整與模擬，耗費許多時間與人力。

為解決上述問題，本研究提出一個以機器學習與演算法為核心的元件反向設計流程。首先，使用 DOE(Design of Experiment)選擇欲蒐集的樣本點，再透過 TCAD 模擬樣本並蒐集為資料集，接著以這些資料集訓練代理模型(Surrogate Model，用於取代 TCAD 的模擬過程)，其中第一層基礎模型為 ANN、Random Forest、SVR、Ridge Regression、第二層模型為 XGBoost。目前，我們的代理模型，對於電性參數 Ron、BV、Coss 的預測準確率都有達 $R^2 > 0.9$，接著進一步，結合基因演算法(NSGA-II)與訓練好的代理模型，執行多目標優化的元件反向設計，依據電性參數與元件衡量指標(Ron 越小越好、BV 越大越好、Coss 越小越好、BFOM 越大越好、HF-FOM 越大越好) ，反向找出最佳的結構參數 gg_distance、lgf、LM1、LM0FP、LGM 的組合。於此同時，我們也使用主動學習(Active Learning)的方式，以選出對於目前模型架構有高度不確定度的樣本，並在這些高不確度的樣本附近做 DOE，選擇接續蒐集的樣本，以提升代理模型的準確度。

綜合來看，使用 Surrogate Model 結合演算法，能有效減少 TCAD 模擬與人為反覆設計元件結構參數的時間，但是我們的 Surrogate Model 的準確度還有提升的空間，可進一步補充資料或是研究出能在少量資料下就訓練出高準確度模型的方法。

關鍵字—GaN HEMT，TCAD，DOE，Surrogate Model，NSGA-II，Active Learning

1. Introduction

1.1 Background

隨著功率電子元件在電動車、電源供應器等領域的應用越來越廣泛，對於電子元件的效率、耐壓與可靠性要求也越來越高，在這樣的趨勢下，氮化鎵高電子遷移率電晶體(GaN HEMT)因其寬能隙、高飽和電子速率、高擊穿電壓等特性，在功率半導體元件中極具潛力。然而，要設計出兼具低導通電阻(Ron)、高崩潰電壓(BV)及低輸出電容(Coss)的 GaN HEMT 元件，需要考慮多個結構參數，而且這些結構參數(甚至電性參數)間往往互相關聯，導致手動設計、調整結構參數極為困難 [6]。

在傳統流程中，工程師多半依靠 TCAD(Technology Computer Aided Design)模擬工具進行參數掃描與優化，耗費大量人力、計算資源與時間。

1.2 Motivation

近年來有不少論文做將機器學習(Machine Learning, ML)應用於半導體製程、元件及電路的模擬與反向設計的研究，再加上，專題教授的實驗室，已有學長建立好的 GaN HEMT 的 TCAD 模擬檔，故本專題選擇 GaN HEMT 做將機器學習應用於電子元件 GaN HEMT 的研究，以解決以下問題:

1. TCAD 模擬時間成本極高: 以本專題的 GaN HEMT 為例，要完整的模擬好一筆樣本，就至少需要 5~6 個小時，而在 GaN HEMT 有五個結構參數，要進行掃描，假設每個參數只取 5 的值，進行掃描，則共有 5^5 = 3125 筆樣本，完整模擬好至少需 15625 小時，651 天。
2. 尋找用少量資料，建立高準確度模型的方法、流程: 由於第一點的原因，造成資料蒐集困難，如果只是憑人為的直覺蒐集資料，所訓練出的代理模型(Surrogate Model)準確度不夠高也不夠有可信度，因此我們需要盡可能地蒐集具代表性的高價值樣本。
3. 多目標設計優化困難: 在 GaN HEMT 設計中，結構參數關聯性高、維度多，需要同時調整多個結構參數(gg_distance、LM0FP、LM1、LGM、lgf)，每個參數都會影響電性參數(Ron、BV、Coss)，導致手動設計極為困難。

1.3 Purpose

基於上述問題，本研究提出一套結合 DOE(Design of Experiment)、代理模型(Surrogate Model)、基因演算法(NSGA-II)、主動學習(Active Learning)、TCAD 模擬與驗證的流程，以提升元件模擬及設計的效率。期望能大幅縮短 GaN

HEMT 元件的設計時間，並提供一種自動化、準確度高的反向設計、優化策略。

此專題研究的主要目標如下：

1. 尋找合適的 DOE(Design of Experiment)方法初步取樣多維結構參數空間，以獲得初始模擬資料。
2. 建立代理模型(Surrogate Model)，輸入: 結構參數(gg_distance、LM0FP、LM1、LGM、lgf)，輸出: 電性參數(Ron、BV、Coss)，以取代 TCAD 費時的模擬。
3. 結合基因演算法(NSGA-II) 與 代理模型(Surrogate Model) 反向的找出，多目標下最優的結構參數組合，以確實的進行元件的反向設計。
4. 結合主動學習(Active Learning) 與 代裡模型(surrogate model)中的 ANN 模型預測時不確定度高的樣本，並在不確定度高的樣本附近做 DOE，以選擇要補充的資料，提升模型準確度。
5. 最終將 3、4 點的優化結果代入 TCAD 進行模擬、驗證。

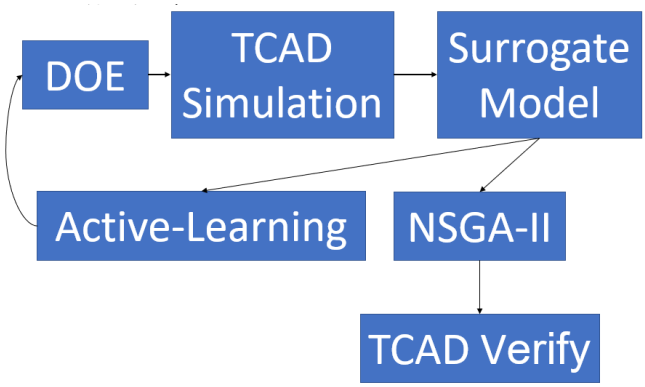


Fig. 1.1 Overall Workflow Architecture

2. Research Methodology

2.1 Description of Device Structure Types

Type 種類	條件 1	條件 2	條件 3	條件 4
Type 1	gg_distance + $Cons_1$ > LM1	gg_distance > LM0FP > $Cons_2$ + LGM + lgf	$Cons_3$ ≥gg_distance≥ $Cons_4$	
Type 2	LM1=0	gg_distance > LM0FP > $Constant_2$ + LGM + lgf	$Cons_3$ ≥gg_distance≥ $Cons_4$	

Type 3	LM1=0 LM0FP=0	gg_distance + $Cons_1$ > LM1	$Cons_3$ ≥gg_distance≥ $Cons_4$	
Type 4	LM1=0 LM0FP=0	gg_distance + $Cons_1$ > LM1	gg_distance > $Cons_2$+ lgf	$Cons_3$ ≥gg_distance≥ $Cons_4$
Type 5	LM0FP=0	gg_distance + $Cons_1$ > LM1	gg_distance > $Cons_2$ + LGM + lgf	$Cons_3$ ≥gg_distance≥ $Cons_4$
Type 7	LM1=0 LGM=0	gg_distance > LM0FP > $Cons_1$ + lgf	$Cons_3$ ≥gg_distance≥ $Cons_4$	
Type 8	LGM=0	gg_distance + $Cons_1$ > LM1	gg_distance > LM0FP > $Cons_2$ + lgf	$Cons_3$ ≥gg_distance≥ $Cons_4$

Table 2.1 Structural Conditions of Each Type

Note: $Cons_1$~$Cons_4$ correspond to $Constant_1$~$Constant_4$,which are the constant values used in the constraint equations.

1、沒有考慮 type_6，是因為 type_6 除了 gg_distance 以外，其他結構都是零，LM1、LM0FP、LGM、lgf 都沒有生成，覺得代表性沒有很好，故沒有花時間蒐集 type_6 的資料。

2、沒有把 type 當作輸入變數(類別變數)，是因為考慮到，type 只是某些結構被設為 0，不同的 type 間在結構上其實還是有一定的關聯性(這個 type 變數，主要是因為寫 tcad 程式碼生成結構時需要，才加入的變數) [5]。

2.2 DOE (Design of Experiments)

在本研究中，我們的目標是於多維結構參數空間中有效取樣，以找到具代表性的樣本作為 TCAD 模擬的輸入，常見的 DOE 方法有 Full Factorial Sampling、Latin Hypercube Sampling (LHS)、D-Optimal Design，然而，這些方法分別以下缺點: 不適合在樣本收集成本高及樣本空間有條件限制時使用、無法處理變數間彼此相依的條件限制、不一定能均勻覆蓋空間，導致未必能反映 GaN HEMT 設計上的結構限制(2.1.2 有說明)，因此最終選擇使用自訂分層的抽樣方法，以確保符合元件結構條件與有價值的取樣，後續將以 type 1 為例進行說明 [7],[8],[9]。

1、問題說明:

type 1 結構參數有 gg_distance > LM0FP > $Cons_2$ + LGM + lgf 的條件，造成「前段變數限制後段變數」的問題，若直接使用 Full Factorial Sampling 或 Latin Hypercube Sampling (LHS)，在 gg_distance 偏小時，後續變數(LM1、LM0FP、LGM、lgf)的可取範圍會被壓縮，使樣本的變數集中於低值的區域，降低各樣本的代表性，也不利於後續代理模型的訓練

2、分段方式說明:

為提升高值區域的探索能力，針對結構參數 gg_distance、LM1、LM0FP 採用自訂的分層分段的抽樣策略，以參數可用範圍的一半作為分割線，分為小值區及大值區，小值區切較少段(5 段)，大值區切較多段(15 段)，使 gg_distance、LM1、LM0FP 較大值的樣本，有更高的選取機率。此外，LGM 與 lgf 由於受限於 LGM + lgf ≤ LM0FP -$Cons_2$，因此採用共同分配的方式進行取樣，先決定 LGM + lgf 可取的範圍，再以隨機比例 α 分配 LGM 與 lgf 的值，以確保兩者在高 gg_distance 時仍能有效率的探索樣本空間。

3、實際抽樣流程說明：

先抽取大量的樣本(不重複)，接著對於每個樣本，依據 gg_distance 的值決定 LM1、LM0FP 的分段上界與分割線，並篩選掉不符合的樣本，接著依據 LM0FP 決定 LGM 、lgf 的上界，再以隨機比例 α 分配 LGM 與 lgf 的值，得到最終生成的樣本，再從中抽取 150 筆，假若不足 150 筆，則再迭代、補抽樣本到 150 筆。此方式確保每個 type 皆能取得覆蓋完整且符合物理限制的代表性設計點。

2.3 Surrogate Model

在本研究中，我們期望能透過機器學習方法，建立可取代 TCAD 模擬的代理模型(surrogate model)，以加速 GaN HEMT 結構的多目標電性參數最佳化流程，由於 GaN 元件的電性(Ron、BV、Coss2)受到多個幾何參數間高度非線性、交互作用影響，單一模型難以穩定學習完整的資料特徵，因此後續使用**兩層式的模型訓練架構**，先透過多個基礎模型(base models)與第二層融合模型(meta model)共同建構高泛化性之預測器，整體架構如下圖所示 [4]。

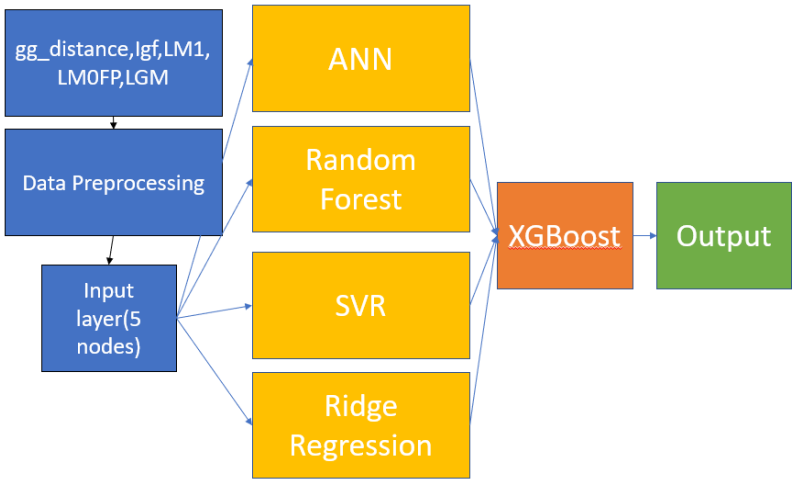


Fig. 2.2 Surrogate Model Structural

2.3.1 輸入特徵與資料前處理

輸入特徵包含五個 GaN HEMT 結構參數 **gg_distance、lgf、LM1、LM0FP、LGM**，資料前處理包含以下內容：

1、**標準化(Standardization):**
採用 z-score 將所有特徵轉換為相同尺度，以避免變數單位、量值影響模型權重。
2、缺失值剔除(Missing Data Removal)
若該筆樣本的 Ron、BV、Coss2 有任一缺失值(TCAD 模擬會有無法收斂的狀況)，則直接移除該樣本(曾嘗試使用 KNN 補值，但因樣本數不夠大以及元件本身非線性的物理特性，使的用補值後資料訓練模型準確度極差偏移物理趨勢，故最終採用直接剔除的策略)
3、切分訓練集、驗證集、測試集資料，比例 70:15:15。

2.3.2 第一層基礎模型(Base Models)
為充分捕捉不同層級的非線性行為，本研究使用四種常見於工程預測之 ML 模型作為第一層基礎模型，以下說明選擇的模型及理由：
1、Artificial Neural Network(ANN)，在資料足夠的情況下，理論上可逼近任意函數，也是四個基礎模型中預測能力最佳的模型。
2、Random Forest 能有效處理變量間的非線性交互作用，對雜訊與不規則資料具高韌性，缺點是易過擬合。
3、Support Vector Regression(SVR) 具備 kernel trick，可建構高維非線性邊界，適合資料量中小且具局部劇變的問題。
4、Ridge Regression 學習變數的線性特徵，確保有學到元件的物理趨勢，提高模型穩定性。

四個模型均以五個結構參數作為輸入，個別預測 Ron、BV、Coss2 三個變數為輸出，第一層基礎模型的總輸出為 $4 \times 3=12$ 維 meta features，作為後續第二層 XGBoost 模型的輸入。

2.3.3 第二層融合模型(Meta Model: XGBoost)
以第一層四個模型對三個電性參數的 12 組預測結果，作為新的 meta features，輸入至第二層的 XGBoost 模型。使用 XGBoost 的理由，能自動學習四個基礎模型在各區域的最佳權重與非線性組合，若某些基礎模型在特定區域預測不佳 XGBoost 能自動降低其重要度，同時在 kaggle 競賽中，XGBoost 也是目前最常用的 meta model。

2.4 Non-dominated Sorting Genetic Algorithm II (NSGA-II)
2.4.1 使用多目標基因演算法(NSGA-II)的理由
本研究的 GaN HEMT 逆向設計涉及三個主要電性參數 Ron、BV、Coss2 及 BFOM、HF-FOM 兩個衍生指標(避免過於演算法優化其中一個電性參數)，屬於多目標、且變數彼此相關的最佳化問題 [1],[2]。例如:
1、減少 gg_distacne 長度，以降低 Ron 但通常會導致 BV 降低。
2、增加 LM0FP、LM1、LGM、lgf 長度可提升 BV 但會造成 Coss2 上升。

3、不同 type 的結構參數具有多個不等式條件。

本研究採用 NSGA-II(Non-dominated Sorting Genetic Algorithm II)進行多目標設計空間搜尋，原因包括：

1、每次迭代搜尋中會得到整條 Pareto 最佳曲線，只要 Surrogate Model 夠準確及優化的目標變數設計得夠好(例如:設其中一個優化目標變數為 BV-10*Ron，就能使 Ron 有較大的權重)，理論上應該能夠找到最佳的元件結構。

2、能同時處理彼此有相關性的電性目標與多個不等式限制條件。

3、可與訓練好的 Surrogate model 快速互動，提高搜尋效率，且每次迭代時保有搜尋的廣度(避免局部最優解)與當前最優解(由交叉變異部分的參數來決定當前最優解的保留比例)。

2.4.2 NSGA-II 整體實作流程

本研究的 NSGA-II 主要流程如下(每個 type 分別執行):

步驟 1 初始化(Initialization):

依各 type 的條件限制，隨機產生 80 筆初始解(父代)。

每一筆透過訓練完成的 surrogate model 預測 Ron、BV、Coss2，再進一步計算 BFOM、HF-FOM。

步驟 2 多目標非支配排序(Non-dominated Sorting):

依據五個目標值(Ron、BV、Coss2、BFOM、HF-FOM)將樣本分類成多個 front。其中 **Front 1** 不被任何樣本支配，為目前最優解，**Front 2 是**僅被 Front 1 支配的解，後續的 front 以此類推

步驟 3 擁擠距離排序(Crowding Distance)

在同個 front 內會進行**擁擠距離排序**，用於衡量樣本在目標空間中的代表性，擁擠距離大，表示位於較稀疏區域，代表性高；擁擠距離小，位於較稠密區，代表性差，此步驟主要用來維持 Pareto 解的多樣性、代表性。

步驟 4 父代選擇(Selection)

從父代中以隨機抽取兩筆，選取 front 較小者，若 front 相同，則選擇擁擠距離大者，重複此程序直到選出 80 筆交配用的父代。

步驟 5 交叉與變異(SBX + Mutation)

使用二進制交叉(Simulated Binary Crossover, SBX)產生新解，使新解在父代的解附近探索，隨後引入小幅度變異，避免演算法只得到局部最優解，最終生成 **80 筆子代**。

步驟 6 合併與存活(Elitism)

將父代(80)與子代(80)合併，共 160 筆，再次進行，非支配排序，擁擠距離排序，選出前 80 筆作為下一代。

重複步驟 2 ~ 6

共 50 代，若進步效率不佳，則會觸發早停機制直接結束當前 type 的基因演算法，最終得到此 type 的 Pareto 解。

2.5 Active Learning

2.5.1 使用主動學習(Active Learning)的理由

透過 MC-Dropout 估計 ANN 模型對於各樣本的不確定度，已達成以下目的：

1、自動找出「模型不確定度大」的樣本，並在這些樣本附近進行 DOE 以補充資料 [3]。

2、能避免用 TCAD 模擬模型已很熟悉的區域，提升優化模的速度。

3、提升 surrogate model 的準確度，避免某些區域殘差值過大，最終在與 NSGA-II 結合，提高反向設計流程的可信度與穩定性。

2.5.2 只對 ANN 使用主動學習(Active Learning)的原因

我們的 Surrogate model 以 ANN、Random Forest、SVR、Ridge Regression 四種模型構成第一層基礎模型，並由 XGBoost 作為第二層模型，然而，基於下述原因，本研究選擇僅對 ANN 進行 active learning，再將補充樣本重新加入 TCAD 模擬。

1、綜合來看 ANN 在三個電性(Ron、BV、Coss2)上的預測最為準確(s)。

2、ANN 可透過 MC-Dropout 自然地產生模型不確定度，適合作為 Active Learning 的不確定度指標。但 Random Forest、SVR、Ridge Regression 皆較難建立有效的模型不確定度評估。

3、第二層模型 XGBoost 屬於 boosting 模型，無法直接做 dropout，也不太適合作為 Active Learning 的不確定度來源。

2.5.3 主動學習流程

使用基於 MC-Dropout 的不確定度估計，流程如下：

1、Dropout

依據之前的 ANN 模型，加入 dropout 的機制，隨機的設某些神經元為零，相當於直接關掉那個神經元的輸出(設 10%)，如此一來即使是完全一樣的輸入變數，每次(目前設 20 次)預測的輸出值(Ron,BV,Coss2)也會因為捨棄的神經元不同而有所不同。

2、不確定度估計

收集多次的預測值(Ron,BV,Coss2)，分別計算這些預測值的標準差，並乘上權重後加總(目前三個變數的權重一樣)，得到不確定度 Uncertainty，不確定度越大，代表模型對這個樣本比較不確定。

每一筆樣本可得到多組預測值：$\widehat{y^{(1)}}, \widehat{y^{(2)}}, \ldots \widehat{y^{(T)}}$

其中電性參數預測值的標準差的加權平均值視為該樣本的不確定度：$U = \omega_1\sigma_{Ron} + \omega_2\sigma_{BV} + \omega_3\sigma_{Coss2}$，其中$\omega_i$為各電性參數的重要度權重(都取 1)。

3、挑選高不確定度樣本

從全部資料中選取不確定度最高的樣本，代表 ANN 對些樣本所在區域較不熟悉或此區域為電性急遽變化區。

4、局部 DOE 採樣

以每一筆高不確定點為中心做局部 DOE 採樣，並加入 type constraint，產生要再次做 TCAD 模擬的結構參數組合。

5、TCAD 模擬、重新訓練 Surrogate Model

將 DOE 產生的新點輸入到 TCAD 模擬，並加入資料集中重新訓練 Surrogate Model。

3 Experimental Results

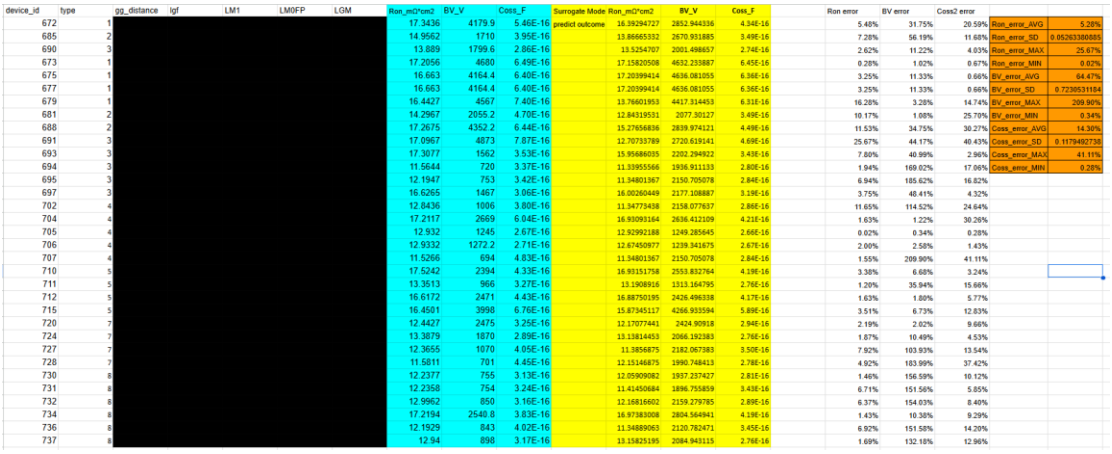
3.1 Surrogate Model Result

Model/電性參數	XGBoost(即為 Surrogate Model)	ANN	SVR	Randon Forest	Ridge Regresion
Ron	0.956	0.927	0.963	0.902	0.994
BV	0.901	0.886	0.895	0.924	0.507
Coss2	0.947	0.749	0.616	0.733	0.731

Table 3.1 The R^2 Values of Each Model for Predicting Each Electrical Parameter

可見 Surrogate Model 對於 Ron、BV、Coss2 的測試資料的R^2都大於 0.9，且相比於任意的基礎模型 ANN、SVR、Random Forest、Ridge Regression 都更加的準確。

3.2 Reverse-Design Validation Results of NSGA-II



device_id	type	gg_distance	lgf	LM1	LM0FP	LGM	Ron_mΩ*cm2	BV_V	Coss_F	Surrogate Mode	Ron_mΩ*cm2	BV_V	Coss_F	Ron error	BV error	Coss2 error		
672	1	[redacted]					17.3436	4179.9	5.46E-16	predict outcome	16.39294727	2852.944336	4.34E-16	5.48%	31.75%	20.59%	Ron_error_AVG	5.28%
685	2						14.9562	1710	3.95E-16		13.86665332	2670.931885	3.49E-16	7.28%	56.19%	11.68%	Ron_error_SD	0.05263380865
690	3						13.889	1799.6	2.86E-16		13.5254707	2001.498657	2.74E-16	2.62%	11.22%	4.03%	Ron_error_MAX	25.67%
673	1						17.2056	4680	6.49E-16		17.15820508	4632.233887	6.45E-16	0.28%	1.02%	0.67%	Ron_error_MIN	0.02%
675	1						16.663	4164.4	6.40E-16		17.20399414	4636.081055	6.36E-16	3.25%	11.33%	0.66%	BV_error_AVG	64.47%
677	1						16.663	4164.4	6.40E-16		17.20399414	4636.081055	6.36E-16	3.25%	11.33%	0.66%	BV_error_SD	0.7230531184
679	1						16.4427	4567	7.40E-16		13.76601953	4417.314453	6.31E-16	16.28%	3.28%	14.74%	BV_error_MAX	209.96%
681	2						14.2967	2055.2	4.70E-16		12.84319531	2077.30127	3.49E-16	10.17%	1.08%	25.70%	BV_error_MIN	0.34%
688	2						17.2675	4352.2	6.44E-16		15.27656836	2839.974121	4.49E-16	11.53%	34.75%	30.27%	Coss_error_AVG	14.30%
691	3						17.0967	4873	7.87E-16		12.70733789	2720.619141	4.69E-16	25.67%	44.17%	40.43%	Coss_error_SD	0.1179492738
693	3						17.3077	1562	3.53E-16		15.95686035	2202.294922	3.43E-16	7.80%	40.99%	2.96%	Coss_error_MAX	41.11%
694	3						11.5644	720	3.37E-16		11.33955566	1936.911133	2.80E-16	1.94%	169.02%	17.06%	Coss_error_MIN	0.28%
695	3						12.1947	753	3.42E-16		11.34801367	2150.705078	2.84E-16	6.94%	185.62%	16.82%		
697	3						16.6265	1467	3.06E-16		16.00260449	2177.108887	3.19E-16	3.75%	48.41%	4.32%		
702	4						12.8436	1006	3.80E-16		11.34773438	2158.077637	2.86E-16	11.65%	114.52%	24.64%		
704	4						17.2157	2669	6.04E-16		16.93093164	2636.412109	4.21E-16	1.63%	1.22%	30.26%		
705	4						12.932	1245	2.67E-16		12.92992188	1249.285645	2.66E-16	0.02%	0.34%	0.28%		
706	4						12.9332	1272.2	2.71E-16		12.67450977	1239.341675	2.67E-16	2.00%	2.58%	1.43%		
707	4						11.5266	694	4.83E-16		11.34801367	2150.705078	2.84E-16	1.55%	209.90%	41.11%		
710	5						17.5242	2394	4.33E-16		16.93151758	2553.832764	4.19E-16	3.38%	6.68%	3.24%		
711	5						13.3513	966	3.27E-16		13.1908916	1313.164795	2.76E-16	1.20%	35.94%	15.66%		
712	5						16.6172	2471	4.43E-16		16.88750195	2426.496338	4.17E-16	1.63%	1.80%	5.77%		
715	5						16.4501	3998	6.76E-16		15.87345117	4266.933594	5.89E-16	3.51%	6.73%	12.83%		
720	7						12.4427	2475	3.25E-16		12.17077441	2424.90918	2.94E-16	2.19%	2.02%	9.66%		
724	7						13.3879	1870	2.89E-16		13.13814453	2066.192383	2.76E-16	1.87%	10.49%	4.53%		
727	7						12.3655	1070	4.05E-16		11.5856875	2182.067383	3.50E-16	7.92%	103.93%	13.54%		
728	7						11.5811	701	4.45E-16		12.15146875	1990.748413	2.78E-16	4.92%	183.99%	37.42%		
730	8						12.2377	755	3.13E-16		12.05909082	1937.237427	2.81E-16	1.46%	156.59%	10.12%		
731	8						12.2358	754	3.24E-16		11.41450684	1896.755859	3.43E-16	6.71%	151.56%	5.85%		
732	8						12.9962	850	3.16E-16		12.16816602	2159.279785	2.89E-16	6.37%	154.03%	8.40%		
734	8						17.2194	2540.8	3.83E-16		16.97383008	2804.564941	4.19E-16	1.43%	10.38%	9.29%		
736	8						12.1929	843	4.02E-16		11.34889063	2120.782471	3.45E-16	6.92%	151.58%	14.20%		
737	8						12.94	898	3.17E-16		13.13825195	2084.943115	2.76E-16	1.69%	132.18%	12.96%		

Fig. 3.6 NSGA-II with Surrogate Model Result

此表格是把 NSGA-II 找出的 pareto 解的預測值(黃色螢光筆)，與 TCAD 模擬值(藍色螢光筆)的比較結果，橘色螢光筆是三個電性參數預測誤差的平均值(AVG)與標準差(SD)，此表可觀察到以下幾點:

1、Ron 誤差平均為 5.28%、最大為 25.67%， BV 誤差平均為 64.47%、最大為 209%，Coss 誤差平均為 14.30%、最大為 41.11%。從表格中可以看到 Ron、Coss 的誤差主要集中在 10%~30%，具有一定的參考性，但 BV 的誤差在部分樣本會高達 100、200%導致幾乎沒有參考性，是之後需要優先改進的目標參數。

2、雖然從之前的 Surrogate Model 的測試集預測結果，以及殘差分析的圖表，看起來預測誤差沒有很大，但模型對於 NSGA-II 找出的 pareto 解的預測誤差卻極大，推測可能是 NSGA-II 探索的區域，與我們 DOE、Active Learning 蒐集資料的範圍有所差距，同時我們數據處理後的有效資料只有 89 筆，導致模型可能還有不穩定的地方。

4.Conclusion and Future Work

此專題實作過程中，我們成功從無到有，摸索出一套用機器學習結合演算法(NSGA-II)反向設計元件的流程架構，內容包含，使用 DOE 選擇進行 TCAD 模擬的樣本，用蒐集的資料訓練 Surrogate Model，將此 Surrogate Model 結合 NSGA-II，以進行元件結構的反向設計，同時使用 Active Learning 的方式選擇要補充的樣本，以進一步提升 Surrogate Model 的準確度。

從前述的實驗方法與結果來看，我們做到以下幾件事情:

1、我們自訂的 DOE 抽樣方法，能有效的在多個不等式條件的情況下，找出高代表性的樣本。

2、Surrogate Model 能快速的依據輸入的結構參數，預測對應的電性參數，約 10 分鐘，就可以預測 50 多筆樣本，若用 TCAD，則模擬一筆資料就需要至少 5、6 個小時，而且容易出現無法收斂的情況。

3、將 Surrogate Model 結合 NSGA-II，能快速設計元件的結構，若單純由人為手動的設計一個元件，往往需要反覆掃描，元件的結構參數，以找出在多個目標最優的元件結構，極為耗費時間與人力。

4、有探索出使用 Active Learning 的方法，能在現有資料下，進一步，收集高代表的樣本，進而快速的迭代，提升 Surrogate Model 的準確度。

5、我們反向設計流程，得到的 pareto 解的，Ron 誤差平均為 5.28%、最大為 25.67%，BV 誤差平均為 64.47%、最大為 209%，Coss 誤差平均為 14.30%、最大為 41.11%。

透過這樣的流程，期望能在保持元件性能的同時，大幅縮短 GaN HEMT 元件的設計時間，並提供一種自動化、準確度高的反向設計、優化策略，並期望此流程，能用於未來的元件設計，減輕工程師的設計負擔。

而後續的延伸、改進方向包含:

1、重新設計 DOE 的取樣方式，可以改成做四個分隔線，再分段的更細緻，也可以針對各個 Type 的元件，分別設計 DOE，因為各 Type 的條件數不太一樣。
2、進一步實作整個 Surrogate Model 或是其他基礎模型的 Active Learing，並且多迭代幾次 Active Learning 反覆蒐集資料，也可直接挑目前 pareto 解有較大電性預測誤差的樣本，附近做 DOE 選擇進一步蒐集的資料，以降低模型在樣本空間中的不確定度
3、可在按業界的實際需求設計 NSGA-II 的目標變數，用類似W_1*(-Ron)+W_2*BV+W_3*(-Coss2) 的式子作為目標變數，依需求調整W_1、W_2、W_3權重的值，例如若較在意 Ron，就可設成W_1:W_2:W_3 為 5:1:1。
4、輸入變數可進一步考慮製程中 doping 的濃度
5、此流程也可以套用於其他複雜元件的設計。

5. Reference

[1] N. Yee, A. Lu, Y. Wang, M. Porter, Y. Zhang and H. Y. Wong, "Rapid Inverse Design of GaN-on-GaN Diode with Guard Ring Termination for BV and (VFQ)−1 Co-Optimization," 2023 35th International Symposium on Power Semiconductor Devices and ICs (ISPSD), Hong Kong, 2023, pp. 143-146, doi: 10.1109/ISPSD57135.2023.10147511. keywords: {Structural rings;Integrated circuits;Breakdown voltage;Smoothing methods;Electric breakdown;Semiconductor diodes;Power semiconductor devices;Breakdown Voltage;Gallium Nitride (GaN);Machine Learning;Technology Computer-Aided Design (TCAD);Pareto Front}

[2] K. Deb, A. Pratap, S. Agarwal and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," in IEEE Transactions on Evolutionary Computation, vol. 6, no. 2, pp. 182-197, April 2002, doi: 10.1109/4235.996017.
keywords: {Genetic algorithms;Sorting;Computational complexity;Evolutionary computation;Computational modeling;Testing;Decision making;Associate members;Diversity reception;Constraint optimization},

[3] Rawat TS, Chang CY, Feng YW, Chen S, Shen CH, Shieh JM, Lin AS. Meta-Learned and TCAD-Assisted Sampling in Semiconductor Laser Annealing. ACS Omega. 2022 Dec 22;8(1):737-746. doi: 10.1021/acsomega.2c06000. PMID: 36643440; PMCID: PMC9836362.

[4] Ridge Regression、SVR、Random Forest、ANN、XGBoost、交叉驗證、殘差分析是從機器學習(孫民教授)、統計學習(鄭又仁教授)及統計資料分析(黃文瀚教授)課程中的學到的。

[5] TCAD 的模擬檔案是實驗室學長(Vito Argono)提供的。

[6] U. K. Mishra, P. Parikh and Yi-Feng Wu, "AlGaN/GaN HEMTs-an overview of device operation and applications," in Proceedings of the IEEE, vol. 90, no. 6, pp. 1022-1031, June 2002, doi: 10.1109/JPROC.2002.1021567.

[7] https://youtu.be/Zs5zvOv_XrI?si=nUt-ydQAXnX79s0i (D-optimal design)

[8] https://www.cnblogs.com/jiangkejie/p/15216657.html (LSH)

[9] https://www.iem.yuntech.edu.tw/lab/qre/public_html/Courses/1/AQM-1/files/CH3%20%E5%85%A8%E5%9B%A0%E5%AD%90%E8%A8%AD%E8%A8%88.pdf (Full Factorial Design)

6. Review and Reflections

在將近一年的專題研究中，從寒假閱讀教授提供的 Paper、補足機器學習與演算法知識，到三下學習 TCAD 操作、設計 DOE 以選出具代表性的樣本，再到暑假開始建立 surrogate model，完整進行了「遇到問題→尋找解決方法→驗證→修正→再嘗試、優化」的研究流程。這段過程讓我深刻體會到工程研究的複雜度遠超課堂上的作業。

專題中期，發現 TCAD 完整模擬一筆資料需至少五到六小時，第一次真正感受到在元件研究中建立好資料集需要極大的時間成本，也迫使我們尋找能取得高代表性樣本的方法，最後由於元件結構的限制，我們沒有使用傳統的 DOE，而是自行設計分層與分段的抽樣策略。

在暑假到四上前半的後期階段，我們正式進入 surrogate model 與 NSGA-II 結合反向設計元件的研究。一開始 ANN 模型的預測能力遠不如預期，R^2甚至不到 0.5。接著嘗試了，各種改善方式，從多頭 ANN、到使用混合模型的方法(四個基礎模型+XGBoost)，逐步把每個電性參數 R^2都提升到 0.9 以上。此外，在四上時想進一步提升資料量，而轉而尋找、使用 Active Learning 的方式，選擇要補充的資料(因為重複地做之前的 DOE 得到的樣本代表性不會太好)，但只來得及迭代一次 Active Learning。上述經驗讓我們體會到，在樣本蒐集成本高的情況下，無法像一般的課堂作業改變幾個模型與參數，就能提升預測的準確度。

最後，感謝黃敬源老師實驗室給的指導與資源，特別感謝學長 Vito Argono 將近一年的帶領與協助，在專題期間陪我們共同面對各種問題，教導我們進行 TCAD 的模擬與處理模擬時出現的各種 bug，這段專題研究經歷不僅讓我們累積了查閱文獻、實作的經驗，也提升了我們面對碩班研究需要的能力。